ECME Thresholding Methods for Sparse Signal Reconstruction

Kun Qiu and *Aleksandar Dogandžić*
ECpE Department, Iowa State University
3119 Coover Hall, Ames, IA 50011
email: {kqiu,ald}@iastate.edu

Abstract

We propose a probabilistic framework for interpreting and developing hard thresholding sparse signal reconstruction methods and present several new algorithms based on this framework. The measurements follow an underdetermined linear model, where the regression-coefficient vector is the sum of an unknown deterministic sparse signal component and a zero-mean white Gaussian component with an unknown variance. We first derive an *expectation-conditional maximization either (ECME) iteration* that guarantees convergence to a local maximum of the likelihood function of the unknown parameters for a given signal sparsity level. To analyze the reconstruction accuracy, we introduce the *minimum sparse subspace quotient (SSQ)*, a more flexible measure of the sampling operator than the well-established restricted isometry property (RIP). We prove that, if the minimum SSQ is sufficiently large, ECME achieves perfect or near-optimal recovery of sparse or approximately sparse signals, respectively. We also propose a *double overrelaxation (DORE)* thresholding scheme for accelerating the ECME iteration. If the signal sparsity level is *unknown*, we introduce an *unconstrained sparsity selection (USS)* criterion for its selection and show that, under certain conditions, applying this criterion is equivalent to finding the sparsest solution of the underlying underdetermined linear system. Finally, we present our *automatic double overrelaxation (ADORE)* thresholding method that utilizes the USS criterion to select the signal sparsity level. We apply the proposed schemes to reconstruct sparse and approximately sparse signals from tomographic projections and compressive samples.

Index Terms

Expectation-conditional maximization either (ECME) algorithm, iterative hard thresholding, sparse signal reconstruction, sparse subspace quotient, unconstrained sparsity selection, overrelaxation.

I. INTRODUCTION

Sparsity is an important concept in modern signal processing. Sparse signal processing methods have been developed and applied to biomagnetic and magnetic resonance imaging, spectral estimation, wireless sensing, and compressive sampling, see [1]–[8] and references therein. For noiseless measurements, the major sparse signal reconstruction task is finding the sparsest solution of an underdetermined linear system $\boldsymbol{y} = H\,\boldsymbol{s}$ (see e.g. [8, eq. (2)]):

$$(\mathrm{P}_0): \qquad \min_{\boldsymbol{s}} \|\boldsymbol{s}\|_{\ell_0} \quad \text{subject to } \boldsymbol{y} = H\,\boldsymbol{s} \tag{1.1}$$

where $\boldsymbol{y}$ is an $N \times 1$ measurement vector, H is a known $N \times m$ full-rank *sensing matrix* with $N \leq m$, $\boldsymbol{s}$ is an $m \times 1$ unknown *signal vector*, and $\|\boldsymbol{s}\|_{\ell_0}$ counts the number of nonzero elements in the signal vector $\boldsymbol{s}$. The (P_0) problem requires combinatorial search and is known to be NP-hard [9].

A number of tractable approaches have been proposed to find sparse solutions to underdetermined systems. They can be roughly divided into three groups: convex relaxation, greedy pursuit, and probabilistic

methods. Convex methods replace the ℓ_0-norm penalty with the ℓ_1-norm penalty and solve the resulting convex optimization problem. Basis pursuit (BP) directly substitutes ℓ_0 with ℓ_1 in the (P_0) problem, see [10]. To combat measurement noise and accommodate for approximately sparse signals, several methods with various optimization objectives have been suggested, e.g. basis pursuit denoising (BPDN) [11], [10] and Dantzig selector [12]. The gradient projection for sparse reconstruction (GPSR) algorithm in [13] solves the unconstrained version of the BPDN problem in a computationally efficient manner. Greedy pursuit methods approximate the (P_0) solution in an iterative manner by making locally optimal choices. Orthogonal matching pursuit (OMP) [14], [15], [16], compressive sampling matching pursuit (CoSaMP) [17], and iterative thresholding schemes [18]–[21] belong to this category. Probabilistic methods utilize full probabilistic models and statistical inference tools to solve the sparse signal reconstruction problem. Examples of the methods in this group are: sparse Bayesian learning (SBL) [22], Bayesian compressive sensing (BCS) [23] and expansion-compression variance-component based method (ExCoV) [24]. Most existing sparse signal reconstruction schemes require *tuning* [25], where the reconstruction performance depends crucially on the choice of the tuning parameters.

Iterative hard thresholding (IHT) and normalized iterative hard thresholding (NIHT) algorithms in [19]–[21] (see also [18]) have attracted significant attention due to their low computation and memory requirements and theoretical and empirical evidence of good reconstruction performance. The IHT and NIHT methods require only matrix-vector multiplications and *do not* involve matrix-matrix products, matrix inversions, or solving linear systems of equations. The memory needed to implement IHT and NIHT is just $\mathcal{O}(Nm)$, and can be further reduced to $\mathcal{O}(m)$ if the sensing operator H is realized in a function-handle form. However, the IHT and NIHT methods

- converge slowly, demanding a fairly large number of iterations,
- require the knowledge of the signal sparsity level, which is a tuning parameter, and
- are sensitive to scaling of the sensing matrix (IHT) or require elaborate adjustments in each iteration to compensate for the scaling problem (NIHT).

IHT and NIHT guarantee good recovery of the underlying sparse signal if the sensing matrix satisfies the *restricted isometry property (RIP)* and a modified non-symmetric RIP; see [20] and [21], respectively. The restricted isometry property was introduced in [7] to measure how well sparse vectors preserve their magnitudes after being transformed by the sensing matrix H. To preserve this magnitude for a sparsity level r, any r columns of H must be approximately orthonormal, which corresponds to H having a small *restricted isometry constant (RIC)*. We refer to this requirement as the *RIP condition*. (See Section IV for the definition of the RIC for sparsity level r, statement of the corresponding RIP condition, and further

discussion.) Besides being used to analyze the IHT and NIHT schemes, the RIP condition is a common ingredient of reconstruction performance analyses of many sparse reconstruction methods, e.g. convex methods [7], [11], [12] and COSAMP [17]. However,

- the RIP condition is quite restrictive: a simple linear transform or even a scaling of H by a constant can easily break the equilibrium required by RIP.

The contribution of this paper is four-fold.

1. Probabilistic model. We propose a probabilistic framework for generalizing iterative hard thresholding (IHT) algorithms and interpreting them as *expectation-conditional maximization either* (ECME) iterations, see also [26]. If the rows of the sensing matrix H are orthonormal, the signal update of the ECME iteration is equivalent to one IHT step. Note that IHT is a greedy pursuit scheme whereas ECME is a probabilistic scheme; hence, our framework blurs the boundary between the two categories.

2. Analysis. We prove that our ECME iteration monotonically converges to a fixed point corresponding to a local maximum of the marginal likelihood function under our probabilistic model. The conditions that we use in this convergence analysis are *invariant* to invertible linear transforms of either the rows or the columns of H, which indicates that the convergence of our ECME iteration is robust to linear transforms of H. Such a convergence robustness to linear transforms and scaling of H is in contrast to the IHT convergence analysis in [19, Theorem 4] that requires the spectral norm of H to be strictly less than one.

We also provide perfect and near-optimal guarantees for the recovery of sparse and approximately sparse signals, respectively. Our signal recovery analysis *does not* rely on the common assumption that H has a sufficiently small RIC; rather, we introduce new measures of H useful for reconstruction analysis: the r-sparse subspace quotient (r-SSQ) and minimum r-SSQ. The minimum r-SSQ measures how well sparse vectors with sparsity level r preserve their magnitudes after being projected onto the row space of H, see Section IV. Unlike the RIC, the minimum r-SSQ is *invariant* to invertible linear transforms of the rows of H. We prove that, if the minimum $2r$-SSQ of the sensing matrix is larger than 0.5, our ECME algorithm for sparsity level r

- *perfectly* recovers the true r-sparse signal from noiseless measurements and
- estimates the best r-term approximation of an arbitrary *non-sparse* signal from *noisy* measurements within a bounded error.

Due to the row transform invariance of the minimum r-SSQ, our reconstruction analysis allows for sensing matrices that violate the RIP condition: the columns of the sensing matrices can have arbitrary magnitudes and be highly correlated. Therefore, our results *widen* the scope of sensing matrices that allow perfect or satisfactory sparse reconstruction performance via tractable algorithms.

3. Convergence acceleration. We develop a *double overrelaxation (DORE)* thresholding method that interleaves two overrelaxation steps with ECME steps, see also [27]. DORE significantly *accelerates* the convergence of the ECME algorithm (and therefore of the IHT method as well, which is its special case). The line searches in the overrelaxation steps have *closed-form* solutions, making these steps computationally efficient. The theoretical convergence and reconstruction properties of ECME in **2** (above) apply to the DORE method as well.

4. Signal sparsity level selection. Finally, we propose an *automatic double overrelaxation (ADORE)* thresholding method that *does not* require the knowledge of the signal sparsity level. To *automatically* select the sparsity level (i.e. estimate it from the data), we introduce an *unconstrained sparsity selection (USS)* model selection criterion. We prove that, under certain mild conditions, the unconstrained criterion USS is *equivalent* to the constrained (P_0) problem (1.1). ADORE combines the USS criterion and DORE iteration and applies a golden-section search to maximize the USS objective function.

In Section II, we introduce our two-stage hierarchical probabilistic model and the ECME thresholding algorithm (Section II-A). Our convergence and near-optimal reconstruction analyses of the ECME iteration are presented in Sections III and IV, respectively. In Section V, we describe the DORE thresholding method for accelerating the convergence of the ECME iteration. In Section VI, we introduce the USS criterion and our ADORE thresholding scheme (Section VI-A). In Section VII, we compare the performances of the proposed and existing large-scale sparse reconstruction methods via numerical experiments. Concluding remarks are given in Section VIII.

A. Notation and Terminology

We introduce the notation used in this paper:

- $\mathcal{N}(\boldsymbol{y}\,;\,\boldsymbol{\mu}, \Sigma)$ denotes the multivariate probability density function (pdf) of a real-valued Gaussian random vector $\boldsymbol{y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ;
- $|\cdot|$, $\|\cdot\|_{\ell_p}$, $\det(\cdot)$, "^{T}" denote the absolute value, ℓ_p norm, determinant, and transpose, respectively;
- the smallest integer larger than or equal to a real number x is $\lceil x \rceil$;
- I_n, $\mathbf{0}_{n \times 1}$, and $0_{n \times m}$ are the identity matrix of size n, the $n \times 1$ vector of zeros, and the $n \times m$ matrix of zeros, respectively;
- $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ are the minimum and maximum eigenvalues of a real-valued symmetric square matrix X;
- $\operatorname{spark}(H)$ is the smallest number of linearly dependent columns of a matrix H [8];
- H_A denotes the *restriction* of the matrix H to the index set A, e.g. if $A = \{1, 2, 5\}$, then $H_A = [\boldsymbol{h}_1 \, \boldsymbol{h}_2 \, \boldsymbol{h}_5]$, where $\boldsymbol{h}_i$ is the ith column of H;

- s_A is the restriction of a column vector s to the index set A, e.g. if $A = \{1, 2, 5\}$, then $s_A = [s_1, s_2, s_5]^T$, where s_i is the ith element of s;

- $\dim(A)$ denotes the size of a set A;

- $\operatorname{supp}(x)$ returns the support set of a vector x, i.e. the index set corresponding to the nonzero elements of x, e.g. $\operatorname{supp}([0, 1, -5, 0, 3, 0]^T) = \{2, 3, 5\}$;

- the thresholding operator $\mathcal{T}_r(x)$ keeps the r largest-magnitude elements of a vector x intact and sets the rest to zero, e.g. $\mathcal{T}_2([0, 1, -5, 0, 3, 0]^T) = [0, 0, -5, 0, 3, 0]^T$.

We refer to an $N \times m$ sensing matrix H as *proper* if it has full rank and

$$N \leq m \tag{1.2}$$

which implies that the rank of H is equal to N. Throughout this paper, we assume that sensing matrices H are proper, which is satisfied in almost all practical sparse signal reconstruction scenarios.

II. Probabilistic Model and the ECME Algorithm

We model a $N \times 1$ real-valued measurement vector y as

$$y = H\,z \tag{2.1a}$$

where H is an $N \times m$ real-valued proper sensing matrix, z is an $m \times 1$ multivariate Gaussian vector with pdf

$$p_{z|\theta}(z\,|\,\theta) = \mathcal{N}(z\,;\,s, \sigma^2\,I_m) \tag{2.1b}$$

$s = [s_1, s_2, \ldots, s_m]^T$ is an *unknown* $m \times 1$ real-valued sparse signal vector containing *at most* r nonzero elements ($r \leq m$), and σ^2 is an unknown *variance-component parameter*; we refer to r as the *sparsity level* of the signal and to the signal s as being r-*sparse*. Note that $\|s\|_{\ell_0} = \dim(\operatorname{supp}(s))$ counts the number of nonzero elements in s; we refer to $\|s\|_{\ell_0}$ as the *support size* of s. Therefore, the support size $\|s\|_{\ell_0}$ of the r-sparse vector s is less than or equal to the sparsity level r. The set of unknown parameters is

$$\theta = (s, \sigma^2) \in \Theta_r \tag{2.2}$$

with the parameter space

$$\Theta_r = \mathcal{S}_r \times [0, +\infty) \tag{2.3a}$$

where

$$\mathcal{S}_r = \{s \in \mathcal{R}^m : \|s\|_{\ell_0} \leq r\} \tag{2.3b}$$

is the sparse signal parameter space. The marginal likelihood function of $\boldsymbol{\theta}$ is obtained by *integrating $\boldsymbol{z}$ out* [see (2.1)]:

$$p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}\,;\,H\,\boldsymbol{s},\sigma^2\,H\,H^T) \tag{2.4a}$$

where the fact that H is a proper sensing matrix ensures that $H\,H^T$ is invertible and, consequently, that the pdf (2.4a) exists. For a given sparsity level r, the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(r) = \big(\widehat{\boldsymbol{s}}_{\mathrm{ML}}(r),\widehat{\sigma}^2_{\mathrm{ML}}(r)\big) = \arg\max_{\boldsymbol{\theta}\in\Theta_r} p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}\big(\boldsymbol{y}\,|\,\boldsymbol{\theta}\big). \tag{2.4b}$$

For any fixed $\boldsymbol{s}$, the marginal likelihood (2.4a) is maximized by

$$\widehat{\sigma}^2(\boldsymbol{s}) = (\boldsymbol{y} - H\,\boldsymbol{s})^T\,(H\,H^T)^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s})\,/\,N. \tag{2.5}$$

Therefore, maximizing (2.4a) with respect to $\boldsymbol{\theta}$ is equivalent to first maximizing the *concentrated likelihood function*

$$p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\boldsymbol{s},\widehat{\sigma}^2(\boldsymbol{s})) = \frac{1}{\sqrt{\det(2\,\pi\,H\,H^T)}}\,[\widehat{\sigma}^2(\boldsymbol{s})]^{-0.5\,N}\,\exp(-0.5\,N) \tag{2.6}$$

with respect to $\boldsymbol{s}\in\mathcal{S}_r$, yielding $\widehat{\boldsymbol{s}}_{\mathrm{ML}}(r)$, and then determining the ML estimate of σ^2 by substituting $\widehat{\boldsymbol{s}}_{\mathrm{ML}}(r)$ into (2.5). Obtaining the exact ML estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(r)$ in (2.4b) requires a combinatorial search and is therefore infeasible in practice. We now present a computationally feasible iterative approach that aims at maximizing (2.4a) with respect to $\boldsymbol{\theta}\in\Theta_r$ and circumvents the combinatorial search.

A. ECME Algorithm For Known Sparsity Level r

We treat $\boldsymbol{z}$ as the *missing (unobserved) data* and present an ECME algorithm for approximately finding the ML estimate in (2.4b), assuming a fixed sparsity level r. Since the sparsity level r is assumed known, we simplify the notation and omit the dependence of the estimates of $\boldsymbol{\theta}$ on r in this section and in Appendix A. An ECME algorithm maximizes *either* the expected complete-data log-likelihood function (where the expectation is computed with respect to the conditional distribution of the unobserved data given the observed measurements) *or* the actual observed-data log-likelihood, see [31, Ch. 5.7].

Assume that the parameter estimate $\boldsymbol{\theta}^{(p)} = \big(\boldsymbol{s}^{(p)},(\sigma^2)^{(p)}\big)$ is available, where p denotes the iteration index. *Iteration $p+1$* proceeds as (see Appendix A for its derivation):

- update the sparse signal estimate using the expectation-maximization (EM) step, i.e. the expectation (E) step:

$$\boldsymbol{z}^{(p+1)} = \mathrm{E}_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}^{(p)}] = \boldsymbol{s}^{(p)} + H^T\,\big(H\,H^T\big)^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s}^{(p)}) \tag{2.7a}$$

followed by the maximization (M) step, which simplifies to

$$\boldsymbol{s}^{(p+1)} = \arg\min_{\boldsymbol{s}\in\mathcal{S}_r} \|\boldsymbol{z}^{(p+1)} - \boldsymbol{s}\|_{\ell_2}^2 = \mathcal{T}_r\big(\boldsymbol{z}^{(p+1)}\big) \qquad (2.7\text{b})$$

and

- update the variance component estimate using the following conditional maximization (CM) step:

$$(\sigma^2)^{(p+1)} = (\boldsymbol{y} - H\,\boldsymbol{s}^{(p+1)})^T\,(H\,H^T)^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s}^{(p+1)})\big/N \qquad (2.7\text{c})$$

obtained by maximizing the marginal likelihood (2.4a) with respect to σ^2 for a fixed $\boldsymbol{s} = \boldsymbol{s}^{(p+1)}$, see (2.5).

In (2.7a), $\mathrm{E}_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}]$ denotes the mean of the pdf $p_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta})$, which is the Bayesian minimum mean-square error (MMSE) estimate of $\boldsymbol{z}$ for *known* $\boldsymbol{\theta}$ [33, Sec. 11.4]. Note that $(H\,H^T)^{-1}$ can be pre-computed before the iteration starts or well approximated by a diagonal matrix; hence, our ECME iteration *does not* require matrix inversions. See Section V-A for detailed discussion on the complexity of the ECME method. If the rows of the sensing matrix H are orthonormal:

$$H\,H^T = I_N \qquad (2.8)$$

then the EM step in (2.7a)–(2.7b) is equivalent to one *iterative hard-thresholding (IHT) step* in [20, eq. (10)].

The above ECME algorithm does not satisfy the general regularity conditions assumed in standard convergence analysis of the EM-type algorithms in e.g. [31] and [32, Theorem 2]. In particular,

- the complete-data and conditional unobserved data given the observed data distributions $p_{\boldsymbol{z},\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{y}\,|\,\boldsymbol{\theta})$ and $p_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta})$ are both degenerate, see (2.1a) and Appendix A;
- the parameter space Θ_r is non-convex and its interior is empty;
- in Θ_r, the partial derivatives of the marginal likelihood (2.4a) with respect to the components of $\boldsymbol{s}$ do not exist for most directions.

Therefore, we establish the convergence of our ECME iteration afresh in the following section.

III. CONVERGENCE ANALYSIS OF THE ECME ALGORITHM

We now answer the following questions. Does the ECME iteration in Section II-A ensure monotonically non-decreasing marginal likelihood (2.4a), does it converge to a fixed point and, if yes, is this fixed point a local or the global maximum of the marginal likelihood function? How do we define a local maximum in the parameter space Θ_r in (2.3a)? Since the sparsity level r is fixed, we omit the dependence of the estimates of $\boldsymbol{\theta}$ on r in this section and in Appendices B and C that contain the proofs of the results of this section.

Maximizing the concentrated likelihood function (2.6) with respect to $s \in \mathcal{S}_r$ is equivalent to minimizing the weighted squared error

$$\mathcal{E}(s) = N \, \widehat{\sigma}^2(s) = (y - H \, s)^T \, (H \, H^T)^{-1} \, (y - H \, s). \tag{3.1}$$

The following identity holds for all $s \in \mathcal{R}^m$ and $s' \in \mathcal{R}^m$:

$$\mathcal{E}(s) = \mathcal{Q}(s \,|\, s') - \mathcal{H}(s \,|\, s') \tag{3.2a}$$

where

$$\mathcal{Q}(s \,|\, s') \;=\; \|s' + H^T(HH^T)^{-1}(y - Hs') - s\|_{\ell_2}^2 \tag{3.2b}$$

$$\mathcal{H}(s \,|\, s') \;=\; (s - s')^T \,[\, I_m - H^T \, (HH^T)^{-1} \, H \,]\, (s - s'). \tag{3.2c}$$

This identity follows by rewriting (3.2b) as $\mathcal{Q}(s \,|\, s') = \|(I_m - H^T(HH^T)^{-1}H)(s' - s) + H^T(HH^T)^{-1}(y - Hs)\|_{\ell_2}^2$ and expanding the squares. Observe that $\mathcal{H}(s \,|\, s')$ is minimized at $s = s'$.

Denote by $s^{(p)}$ the estimate of s obtained in *Iteration* p of our ECME iteration. When we set $s' = s^{(p)}$, $\mathcal{Q}(s \,|\, s^{(p)}) = \|z^{(p+1)} - s\|_{\ell_2}^2$ becomes exactly the expression that is minimized in the M step (2.7b) and, consequently,

$$\mathcal{Q}(s^{(p+1)} \,|\, s^{(p)}) \leq \mathcal{Q}(s^{(p)} \,|\, s^{(p)}). \tag{3.3a}$$

Since $\mathcal{H}(s \,|\, s^{(p)})$ is minimized at $s = s^{(p)}$, we have

$$\mathcal{H}(s^{(p+1)} \,|\, s^{(p)}) \geq \mathcal{H}(s^{(p)} \,|\, s^{(p)}). \tag{3.3b}$$

Subtracting (3.3a) from (3.3b) and using (3.2a) yields

$$\mathcal{E}(s^{(p+1)}) \leq \mathcal{E}(s^{(p)}) \tag{3.4}$$

and, therefore, our ECME iteration (2.7) ensures a *monotonically non-decreasing* marginal likelihood (2.4a), see also (2.6). Monotonic convergence is also a key general property of the EM-type algorithms [31]. Furthermore, since (3.1) is bounded from below by zero, the sequence $\mathcal{E}(s^{(p)})$ must converge to a limit as the iteration index p grows to infinity.

However, the fact that $\mathcal{E}(s^{(p)})$ converges does not necessarily imply that $s^{(p)}$ converges to a fixed point. The following theorem establishes convergence of the ECME signal iterates $s^{(p)}$.

Theorem 1: Assume that the sparsity level r satisfies

$$r \leq \tfrac{1}{2} \, (m - N) \tag{3.5a}$$

and that the sensing matrix H satisfies the unique representation property (URP) [1] stating that all $N \times N$ submatrices of H are invertible or, equivalently, that

$$\text{spark}(H) = N + 1. \tag{3.5b}$$

Then, the ECME signal iterate $\boldsymbol{s}^{(p)}$ for sparsity level r converges monotonically to its fixed point as the iteration index p grows to infinity.

Proof: See Appendix B. □

Note that (3.5a) is a mild condition. In practice, $N \ll m$ and (3.5a) specifies a large range of sparsity levels r for which the ECME iteration converges to its fixed point.

Theorem 1 guarantees the convergence of our ECME iteration to a fixed point. However, can we guarantee that this fixed point is a local or the global maximum of the marginal log-likelihood function (2.4a)? To answer this question, we first define the local maximum of a function over the parameter space $\mathcal{S}_r$ in (2.3b).

Definition 1: **r-local maximum and minimum.** For a function $f(\boldsymbol{s}) : \mathcal{R}^m \to \mathcal{R}$, a vector $\boldsymbol{s}^\star \in \mathcal{S}_r$ is an r-local maximum point of $f(\boldsymbol{s})$ if there exists a $\delta > 0$, such that, for all $\boldsymbol{s} \in \mathcal{S}_r$ satisfying $\|\boldsymbol{s} - \boldsymbol{s}^\star\|_{\ell_2} < \delta$, we have

$$f(\boldsymbol{s}^\star) \geq f(\boldsymbol{s}).$$

Then, $f(\boldsymbol{s}^\star)$ is the corresponding r-local maximum of $f(\boldsymbol{s})$. We define $\boldsymbol{s}^\star \in \mathcal{S}_r$ and $f(\boldsymbol{s}^\star)$ as an r-local minimum point and the corresponding r-local minimum of $f(\boldsymbol{s})$ if $\boldsymbol{s}^\star$ is an r-local maximum point for the function $-f(\boldsymbol{s})$.

Definition 1 states that an r-sparse vector is a r-local maximum (or minimum) point of a function $f(\boldsymbol{s})$ if, in some small neighborhood, this vector attains the largest (or smallest) function value among all the sparse vectors within that small neighborhood. Fig. 1 illustrates this concept using $\boldsymbol{s} = [s_1, s_2]^T$ (i.e. $m = 2$) and $f(\boldsymbol{s}) = \exp\{-0.5\,[(s_1 + 0.5)^2 + (s_2 - 0.7)^2]\}$. For the sparsity level $r = 1$, the points $\boldsymbol{a} = [-0.5, 0]^T$ and $\boldsymbol{b} = [0, 0.7]^T$ are the only two 1-local maximum points of $f(\boldsymbol{s})$. Observe that $\boldsymbol{a}$ and $\boldsymbol{b}$ are *not* local maximum points of $f(\boldsymbol{s})$ when $\boldsymbol{s}$ is unconstrained in $\mathcal{R}^2$.

The following lemma provides a necessary condition for an r-local maximum or minimum point of a differentiable function.

Lemma 1: If an r-sparse vector $\boldsymbol{s}^\star \in \mathcal{S}_r$ is an r-local maximum or minimum point of a differentiable function $f(\boldsymbol{s}) : \mathcal{R}^m \to \mathcal{R}$, then, for all $i \in \{1, 2, \ldots, m\}$ such that

$$\dim\big(\{i\} \cup \text{supp}(\boldsymbol{s}^\star)\big) \leq r \tag{3.6a}$$

we have

$$\left.\frac{\partial f(\boldsymbol{s})}{\partial s_i}\right|_{\boldsymbol{s}=\boldsymbol{s}^\star} = 0. \tag{3.6b}$$
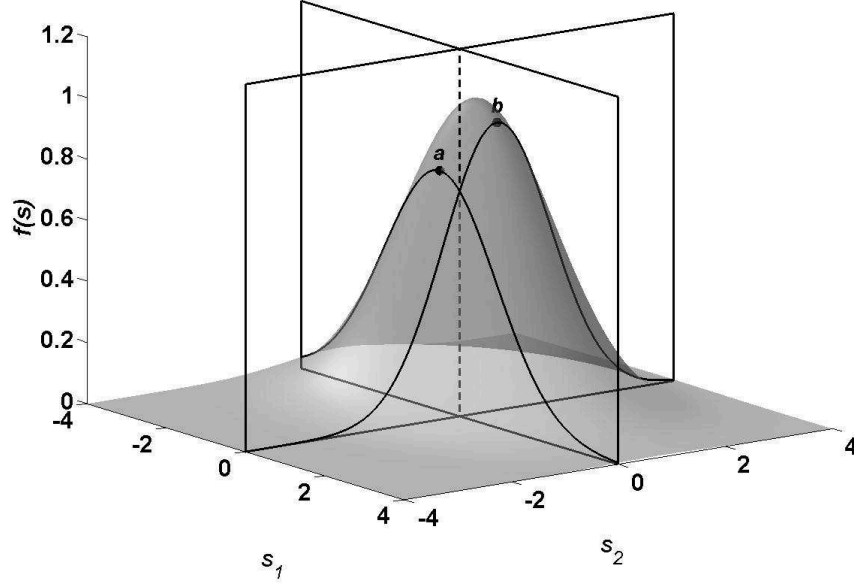
Fig. 1. Function $f(\boldsymbol{s}) = \exp\{-0.5\left[(s_1 + 0.5)^2 + (s_2 - 0.7)^2\right]\}$ with $\boldsymbol{s} = [s_1, s_2]^T$ and two 1-local maxima of $f(\boldsymbol{s})$.

Proof: See Appendix C. □

The condition (3.6a) of Lemma 1 implies that, instead of checking that all partial derivatives of our function are zero (which is required in the standard first-derivative test for finding local maxima and minima), we only need to check its derivatives along a few *allowed* coordinate axes, where the allowed coordinate axes are defined by the property that perturbing along these axes does not violate the sparsity requirement, see (3.6a). If $\boldsymbol{s}^\star$ has exactly r nonzero elements, then i in (3.6a) must belong to $\mathrm{supp}(\boldsymbol{s}^\star)$, and we should only check the r partial derivatives that correspond to the nonzero components of $\boldsymbol{s}^\star$. For example, consider Fig. 1: to determine if $\boldsymbol{a} = [-0.5, 0]^T$ is a 1-local maximum point, we only need to check that the partial derivative of $f(\boldsymbol{s})$ with respect to s_1 is zero at $\boldsymbol{s} = \boldsymbol{a}$; the direction along the s_2 axis is not allowed because the perturbation along this direction violates the sparsity requirement. However, when $\boldsymbol{s}^\star$ has less than r nonzero elements, we must check all partial derivatives, because perturbing along any axis will not exceed the sparsity requirement.

We now provide a sufficient condition for an r-local maximum or minimum point of a twice differentiable function.

Lemma 2: An r-sparse vector $\boldsymbol{s}^\star \in \mathcal{S}_r$ is an r-local maximum or minimum of a twice differentiable function $f(\boldsymbol{s}) : \mathcal{R}^m \to \mathcal{R}$ if

(1)   for all $i \in \{1, 2, \ldots, m\}$ such that $\dim(\{i\} \cup \mathrm{supp}(\boldsymbol{s}^\star)) \leq r$, we have

$$\frac{\partial f(\boldsymbol{s})}{\partial s_i}\bigg|_{\boldsymbol{s}=\boldsymbol{s}^\star} = 0 \tag{3.7}$$

and

(2)   there exists a $\delta > 0$, such that, for all $\boldsymbol{s} \in \mathcal{S}_r$ satisfying $\|\boldsymbol{s} - \boldsymbol{s}^\star\|_{\ell_2} < \delta$, the Hessian matrix

$$\frac{\partial^2 f(\boldsymbol{s})}{\partial \boldsymbol{s}\, \partial \boldsymbol{s}^T}$$

is negative semidefinite (for a maximum) or positive semidefinite (for a minimum).

Proof: See Appendix C.   □

In the example depicted in Fig. 1, both points $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy the two conditions of Lemma 2 and are therefore the r-local maxima of $f(\boldsymbol{s})$. Lemma 2 is useful in developing the following theorem stating that our ECME algorithm actually converges to an r-local maximum point of the concentrated marginal likelihood function (2.6).

Theorem 2: If the sensing matrix H is proper and ECME iteration in Section II-A converges to a fixed point $\boldsymbol{\theta}^\star = (\boldsymbol{s}^\star, (\sigma^2)^\star)$, then $\boldsymbol{s}^\star$ is an r-local maximum point of the concentrated marginal likelihood function (2.6).

Proof: See Appendix C.   □

Based on Theorems 1 and 2, we claim that, if H satisfies the URP condition (3.5b) and for a sufficiently small sparsity level r, the ECME algorithm in Section II-A converges to a fixed point that is an r-local maximum of the concentrated marginal likelihood function (2.6).

The conditions of Theorems 1 and 2 hold even when the sensing matrix H is pre- or post- multiplied by a full rank square matrix. In contrast, the IHT algorithm converges to a local minimum of the squared residual error for a specified sparsity level *only* if H is appropriately scaled. Indeed, Theorem 4 in [19] demands that the spectral norm of the sensing matrix H is less than unity. If the spectral norm condition is violated, the IHT iteration may become unstable and diverge, see [21, Sec. II-D]. To overcome such scaling requirements and ensure convergence for an arbitrary scaled H, a normalized IHT (NIHT) method has been proposed in [21], where a scaling term is introduced to the original hard thresholding step; this term must be monitored and adjusted in each iteration so that it does not exceed a certain threshold (see [21, e.q. 14]); otherwise, the squared residual error [19, eq. (1.6)] is not guaranteed to decrease during the iteration. However, this monitoring and adjustment consume CPU time and typically *slow down* the resulting algorithm, see the numerical examples in Section VII. In contrast, Theorems 1 and 2 assert that the monotonic convergence of our ECME iteration is not affected by the pre- and post-multiplication of H

by any full-rank square matrix of appropriate size, thus removing the need for monitoring and adjustment within the iteration steps.

IV. SPARSE SUBSPACE QUOTIENT AND NEAR-OPTIMAL ECME RECONSTRUCTION

We now study theoretical guarantees for near-optimal ECME reconstruction. We first define the r-sparse subspace quotient (r-SSQ) as a normalized squared magnitude of the projection of an r-sparse signal onto the row space of the sensing matrix. We introduce minimum r-sparse subspace quotient of the sensing matrix as a separability measure for arbitrary r-sparse signals, discuss its properties, compare it with the existing popular measures such as restricted isometry and coherence, and use it to establish a condition for uniqueness of the solution to the (P_0) problem. We then show that, in the absence of noise and if the minimum $2r$-sparse subspace quotient is sufficiently large, our ECME algorithm estimates the true unknown r-sparse signal *perfectly* from the linear measurements. We also give an example of the existence of low-dimensional matrices that satisfy our perfect recovery requirement (Section IV-A). We finally show that, for non-sparse signals and noisy measurements, the ECME iteration for the sparsity level r recovers the best r-term approximation of the true signal within a bounded error. Both the noiseless and noisy reconstruction guarantees hold regardless of the initial estimate of the signal parameters $\boldsymbol{\theta}$ employed by the ECME iteration.

*Definition 2: **r-Sparse Subspace Quotient (r-SSQ) and minimum r-SSQ.*** We define the r-sparse subspace quotient of a nonzero r-sparse vector $\boldsymbol{s}$ of size $m \times 1$ (i.e. $\boldsymbol{s} \in \mathcal{S}_r \backslash \boldsymbol{0}_{m\times1}$) and a proper $N \times m$ sensing matrix H as the ratio of the squared magnitude of the projection of $\boldsymbol{s}$ onto the row space of H and the squared magnitude of $\boldsymbol{s}$:

$$\rho_r(\boldsymbol{s}, H) \triangleq \frac{\|H^T (H\,H^T)^{-1} H\,\boldsymbol{s}\|_{\ell_2}^2}{\|\boldsymbol{s}\|_{\ell_2}^2} = \frac{\boldsymbol{s}^T H^T (H\,H^T)^{-1} H\,\boldsymbol{s}}{\boldsymbol{s}^T\,\boldsymbol{s}}. \tag{4.1a}$$

Define the corresponding minimum r-sparse subspace quotient of the sensing matrix H as

$$\rho_{r,\min}(H) \triangleq \min_{\boldsymbol{s}\in\mathcal{S}_r\backslash\boldsymbol{0}_{m\times1}} \rho_r(\boldsymbol{s}, H). \tag{4.1b}$$

Note that $H^T (HH^T)^{-1} H$ is the projection matrix onto the row space of H and the second equality in (4.1a) follows from the fact that the projection matrix is idempotent.

The following lemma summarizes a few useful properties of r-SSQ and minimum r-SSQ.

Lemma 3: For an $N \times m$ proper sensing matrix H, a nonzero r-sparse vector $\boldsymbol{s}$ of size $m \times 1$, and a sparsity level r satisfying $0 < r \leq m$,

(a) $\rho_r(\boldsymbol{s}, H)$ in (4.1a) can be equivalently defined as

$$\rho_r(\boldsymbol{s}, H) = \frac{\boldsymbol{s}_A^T H_A^T (H\,H^T)^{-1} H_A\,\boldsymbol{s}_A}{\boldsymbol{s}_A^T\,\boldsymbol{s}_A} \tag{4.2a}$$

where $A = \mathrm{supp}(s)$ is the support set of s and $\rho_{r,\min}(H)$ can be determined by the following equivalent optimization:

$$\rho_{r,\min}(H) = \min_{A \subseteq \{1,2,\dots,m\},\, \dim(A)=r} \lambda_{\min}\big(H_A^T \,(H\,H^T)^{-1}\,H_A\big); \tag{4.2b}$$

(b) $\rho_r(s, H)$ and $\rho_{r,\min}(H)$ are invariant to invertible linear transforms of the rows of H, i.e.

$$\rho_r(s, H) = \rho_r(s, G\,H), \quad \rho_{r,\min}(H) = \rho_{r,\min}(G\,H) \tag{4.3}$$

for any full-rank $N \times N$ matrix G;

(c) $\rho_r(s, H)$ and $\rho_{r,\min}(H)$ are bounded as follows:

$$0 \le \rho_{r,\min}(H) \le \rho_r(s, H) \le 1 \tag{4.4}$$

where $\rho_{r,\min}(H)$ attains

- the lower bound 0 when $r > N$ and
- the upper bound 1 when $N = m$;

(d) if and only if H has at least r linearly independent columns, i.e.

$$\mathrm{spark}(H) > r \tag{4.5}$$

the following strict inequality holds:

$$\rho_{r,\min}(H) > 0; \tag{4.6}$$

(e) if $0 < r_1 < r_2$, then

$$\rho_{r_1,\min}(H) \ge \rho_{r_2,\min}(H). \tag{4.7}$$

Proof: See Appendix D. □

We now compare minimum r-SSQ with the commonly used restricted isometry property (RIP) [6], [7], [11], [12], [17], [20], [21] and coherence [8], [15], [16], [18], [37], [38]. The idea behind RIP is to upper-bound deviations of the squared magnitude of $H\,s$ from the squared magnitude of s for arbitrary nonzero r-sparse vectors s; therefore, the following quotient should be close to unity for arbitrary nonzero r-sparse s:

$$\frac{\|H s\|_{\ell_2}^2}{\|s\|_{\ell_2}^2} = \frac{s^T \, H^T \, H \, s}{s^T \, s}. \tag{4.8}$$

The restricted isometry constant (RIC) for sparsity level r can be written as (see [7, e.q. (1.7)]):

$$\gamma_r(H) = \max_{s \in \mathcal{S}_r \setminus \mathbf{0}_{m \times 1}} \left| 1 - \frac{\|H s\|_{\ell_2}^2}{\|s\|_{\ell_2}^2} \right| = \max_{s \in \mathcal{S}_r \setminus \mathbf{0}_{m \times 1}} \left| 1 - \frac{s^T \, H^T \, H \, s}{s^T \, s} \right| \tag{4.9}$$

which quantifies the largest-magnitude deviation of (4.8) from unity. Clearly, the smaller the r-RIC is, the closer to orthonormal any r columns of H are. The assumption that the appropriate RIC is sufficiently small

is key for sparse-signal recovery analyses of the IHT algorithms [20], [21], COSAMP [17], and convex relaxation methods [7], [11], [12]. The coherence measures the largest-magnitude inner product of any two distinct columns of H. The assumption that the coherence is small is a basis for sparse-signal recovery analyses of convex relaxation methods in [38] and of greedy methods (such as OMP) in [15]. However, both the RIP and coherence requirements are somewhat fragile: a simple linear transform or even a scaling of H by a constant can easily break the equilibria required by the RIP or coherence. In comparison, the minimum r-SSQ in (4.1b) measures the smallest normalized squared magnitude of the projection of an r-sparse signal onto the row space of the sensing matrix H. Here, it is the *row space* of H that matters, rather than H itself. Lemma 3 (b) states that the minimum r-SSQ is *invariant* to invertible linear transforms of the rows of H. Therefore, the sensing matrix H can be pre-multiplied by any $N \times N$ full-rank matrix,[1] leading to arbitrary column magnitudes and highly correlated columns, while still keeping the same minimum r-SSQ value. Hence, minimum r-SSQ is a more flexible property of H than RIP and coherence.

We now utilize the minimum SSQ measure to establish a condition under which the solution to the (P_0) problem is unique and leads to exact recovery under the noiseless scenario. A similar problem is considered in [7, Lemma 1.2] and [8, Theorem 2], where such uniqueness and exact recovery conditions have been derived using RIP (4.9) and spark.

Lemma 4: Suppose that we have collected a measurement vector $\boldsymbol{y} = H\boldsymbol{s}^\diamond$ using a proper sensing matrix H, where $\boldsymbol{s}^\diamond$ is a sparse signal vector having exactly $\|\boldsymbol{s}^\diamond\|_{\ell_0} = r^\diamond$ nonzero elements. If the minimum $2r^\diamond$-SSQ of the sensing matrix H is strictly positive:

$$\rho_{2r^\diamond, \min}(H) > 0 \tag{4.10}$$

then the solution to the (P_0) problem (1.1) is unique and coincides with $\boldsymbol{s}^\diamond$.

Proof: See Appendix D. □

Observe that the condition (4.10) implies that the number of measurements N is larger than or equal to twice the support size of the true sparse signal $\boldsymbol{s}^\diamond$, i.e.

$$N \geq 2\,r^\diamond = 2\,\|\boldsymbol{s}^\diamond\|_{\ell_0}. \tag{4.11}$$

Indeed, if $N < 2\,r^\diamond$, $\rho_{2r^\diamond, \min}(H) = 0$ by part (c) of Lemma 3.

Lemma 4 also holds if we replace $r^\diamond$ in the condition (4.10) with any $r > r^\diamond$, which follows from part (e) of Lemma 3: if $\rho_{2r, \min}(H) > 0$ for $r^\diamond < r$, then $\rho_{2r^\diamond, \min}(H) \geq \rho_{2r, \min}(H) > 0$. Therefore, (4.10) is the weakest condition on H among all $r \geq r^\diamond$.

---

[1]Unlike the ECME convergence analysis in Section III, invertible linear transforms of the columns of H are generally not allowed here.

In a nutshell, Lemma 4 states that, for a strictly positive $\rho_{2r,\min}(H)$, any two distinct r-sparse vectors can be distinguished from their projections onto the row space of H, which furthermore guarantees the uniqueness of the (P_0) problem. Note that [7, Lemma 1.2] states that the solution to the (P_0) problem (1.1) is unique and coincides with $\boldsymbol{s}^\diamond$ if the $2r^\diamond$-RIC of the sensing matrix H satisfies

$$\gamma_{2r^\diamond}(H) < 1. \tag{4.12}$$

However, for proper sensing matrices, the condition (4.10) of Lemma 4 is *weaker* than (4.12): (4.12) implies that $\mathrm{spark}(H) > 2\,r^\diamond$ [see (4.9)] and, consequently, (4.10), but not vice versa. For example, the 2×3 sensing matrix

$$H = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{4.13}$$

satisfies the condition (4.10) with $\rho_{2,\min}(H) = 1/3 > 0$, but violates (4.12), since its 2-RIC is $\gamma_2(H) = 1.618 > 1$. Hence, (4.10) *does not* imply (4.12).

We now develop reconstruction performance guarantees for our ECME algorithms that employ the minimum r-SSQ measure.

Theorem 3: **Exact Sparse Signal Reconstruction From Noiseless Samples.** Suppose that we have collected a measurement vector

$$\boldsymbol{y} = H\,\boldsymbol{s}^\diamond \tag{4.14a}$$

where $\boldsymbol{s}^\diamond \in \mathcal{S}_r$ is an r-sparse signal vector, i.e. $\|\boldsymbol{s}^\diamond\|_{\ell_0} \leq r$. If the minimum $2r$-SSQ of the sensing matrix H satisfies

$$\rho_{2r,\min}(H) > 0.5 \tag{4.14b}$$

then the ECME iteration for the sparsity level r in Section II-A converges to the ML estimate of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(r) = (\boldsymbol{s}^\diamond, 0) \tag{4.14c}$$

and therefore recovers the true sparse signal $\boldsymbol{s}^\diamond$ perfectly.

Proof: See Appendix D. □

Theorem 3 shows that, upon convergence and if the minimum $2r$-SSQ of the sensing matrix is sufficiently large, the ECME algorithm recovers the true sparse signal $\boldsymbol{s}^\diamond$ *perfectly* from the noiseless measurements. In this case, the ECME iteration converges to the *global maximum* of the marginal likelihood (2.4a), which is infinitely large since the ML estimate of σ^2 is zero. This global convergence is guaranteed regardless of the initial estimate of $\boldsymbol{\theta}$ used to start the ECME iteration. In addition, by Lemma 4, $\boldsymbol{s}^\diamond$ is also the unique

solution to the (P_0) problem. Therefore, under the conditions of Theorem 3, the ECME algorithm solves the (P_0) problem as well.

Next, we consider a more practical scenario where the true signal $s^\diamond$ is not strictly sparse and the measurements y are corrupted by noise.

Theorem 4: **Near-Optimal Recovery of Non-sparse Signal From Noisy Samples.** Suppose that we have collected a measurement vector

$$y = Hs^\diamond + n \tag{4.15a}$$

where the signal $s^\diamond$ is not necessarily sparse and $n \in \mathcal{R}^N$ is a noise vector. Denote by $s_r^\diamond$ the best r-term ℓ_2-norm approximation to $s^\diamond$, i.e.

$$s_r^\diamond = \arg\min_{s \in \mathcal{S}_r} \|s - s^\diamond\|_{\ell_2} = \mathcal{T}_r(s^\diamond) \tag{4.15b}$$

and by $s^\star$ the r-sparse signal estimate obtained upon convergence of the ECME iteration for the sparsity level r in Section II-A. If the minimum $2r$-SSQ of the sensing matrix H satisfies

$$\rho_{2r,\min}(H) > 0.5 \tag{4.15c}$$

which is the same as the condition (4.14b) in Theorem 3, then

$$\|s^\star - s_r^\diamond\|_{\ell_2} \le 2 \frac{\|s^\diamond - s_r^\diamond\|_{\ell_2} + \|H^T(HH^T)^{-1}n\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)} - \sqrt{1 - \rho_{2r,\min}(H)}}. \tag{4.15d}$$

Proof: See Appendix D. $\qquad\square$

Theorem 4 shows that, for a general (not necessarily sparse) signal $s^\diamond$ and noisy measurements satisfying (4.15a) and sensing matrix satisfying (4.15c), the ECME estimate is close to the best r-term ℓ_2-norm approximation of $s^\diamond$. This result holds regardless of the initial estimate of θ employed by the ECME iteration. Observe that, by (4.4), $\rho_{2r,\min}(H) \le 1$ and therefore the squared roots in (4.15d) are well-defined. Moreover, since (4.15c) holds, the denominator on the right-hand side of (4.15d) is positive and less than or equal to one. When the noise n is zero and signal $s^\diamond$ is r-sparse, the quantities $\|s^\diamond - s_r^\diamond\|_{\ell_2}$ and $\|H^T(HH^T)^{-1}n\|_{\ell_2}$ in (4.15d) are zero and, therefore, $\|s^\star - s^\diamond\|_{\ell_2} = 0$, consistent with Theorem 3.

Performance guarantees similar to those in Theorems 3 and 4 have been developed for other sparse reconstruction methods. However, these results rely on either small RIP constants (see e.g. [7, Theorems 1.3, 1.4], [11, Theorem 1], [12, Theorem 1.1] and [17, Theorem A], [20, Theorems 4, 5], [21, Theorem 4]) or small coherence (see e.g. [38, Theorem 2] and [15, Theorem 3.5]). Therefore, all previous results require that a certain numbers of columns of the sensing matrix H are approximately orthonormal (RIP) or orthogonal (coherence). In contrast, our analysis of the ECME method in Theorems 3 and 4 applies

to the cases where the columns of H are not approximately orthonormal and can be heavily correlated, thus widening the class of sensing matrices for which it is possible to derive reconstruction performance guarantees, see also the discussion after Lemma 3.

A. An Example of a Low-dimensional Matrix Satisfying The Conditions of Theorems 3 and 4

The ongoing search for desirable sensing matrices focuses on small RIP constants and on asymptotic behavior of *large* random matrices, e.g. Gaussian, Bernoulli (with entries equal to 1 and -1), and Fourier (randomly selected rows of the DFT matrix) matrices, see e.g. [7] and [39]. We now show that it is possible to find *low-dimensional* sensing matrices that satisfy the condition $\rho_{2r,\min}(H) > 0.5$ of Theorems 3 and 4.

Consider the 21×32 sensing matrix H comprised of the 21 rows of the 32×32 type-II discrete cosine transform (DCT) matrix (see e.g. [40, Sec. 8.8.2]) with indices

$$2, 3, 4, 5, 7, 9, 10, 12, 13, 14, 16, 18, 20, 21, 22, 24, 27, 29, 30, 31, 32. \tag{4.16}$$

It can be verified by combinatorial search that the minimum 2-SSQ of H meets the condition (4.14b):

$$\rho_{2,\min}(H) = 0.503 > 0.5 \tag{4.17}$$

and, therefore, by Theorem 3, for this H, our ECME iteration *perfectly* recovers any 1-sparse signal s from the 21 noiseless linear measurements given by $y = Hs$. (We have checked and confirmed the validity of this statement via numerical simulations.) However, the 2-RIC of the same 21×32 sensing matrix H is

$$\gamma_2(H) = 0.497. \tag{4.18}$$

which violates the condition required in the theoretical analysis of the IHT algorithm [20, Theorems 4 and 5]. In particular, Theorems 4 and 5 in [20] require $\gamma_3(H) < 1/\sqrt{32} \approx 0.177$ for 1-sparse signals, but here $\gamma_3(H) \geq \gamma_2(H) = 0.497$. We have checked that the above sensing matrix H also violates the condition required in the theoretical analysis of the NIHT algorithm [21, Theorem 4]. Indeed, for 1-sparse signals and the above sensing matrix H, the *non-symmetric restricted isometry* constant in [21] is at least 0.611, which is larger than the upper limit 0.125, see [21, Theorems 4].

Due to the invariance property of Lemma 3 (b), any invertible linear transformation of the rows of H *preserves* the minimum r-SSQ constant. Therefore, upon finding one good sensing matrix H that satisfies $\rho_{2r,\min}(H) > 0.5$, we can construct infinitely many matrices that satisfy this condition.

V. THE DORE ALGORITHM FOR KNOWN r

We now present the DORE thresholding method that accelerates the convergence of our ECME iteration. Since the sparsity level r is assumed known, we omit the dependence of the estimates of θ on r in this section.

Assume that two consecutive estimates of the unknown parameters $\boldsymbol{\theta}^{(p-1)} = (\boldsymbol{s}^{(p-1)}, (\sigma^2)^{(p-1)})$ and $\boldsymbol{\theta}^{(p)} = (\boldsymbol{s}^{(p)}, (\sigma^2)^{(p)})$ are available from the $(p-1)$-th and p-th iterations, respectively. *Iteration $p+1$* proceeds as follows:

1. ECME step. Compute

$$\widehat{\boldsymbol{s}} = \mathcal{T}_r\big(\boldsymbol{s}^{(p)} + H^T(HH^T)^{-1}(\boldsymbol{y} - H\boldsymbol{s}^{(p)})\big) \tag{5.1a}$$

$$\widehat{\sigma}^2 = (\boldsymbol{y} - H\widehat{\boldsymbol{s}})^T (HH^T)^{-1} (\boldsymbol{y} - H\widehat{\boldsymbol{s}})/N \tag{5.1b}$$

and define $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{s}}, \widehat{\sigma}^2)$.

2. First overrelaxation. Compute the linear combination of $\widehat{\boldsymbol{s}}$ and $\boldsymbol{s}^{(p)}$:

$$\bar{\boldsymbol{z}} = \widehat{\boldsymbol{s}} + \alpha_1 (\widehat{\boldsymbol{s}} - \boldsymbol{s}^{(p)}) \tag{5.2a}$$

where the weight

$$\alpha_1 = \frac{(H\widehat{\boldsymbol{s}} - H\boldsymbol{s}^{(p)})^T (HH^T)^{-1} (\boldsymbol{y} - H\widehat{\boldsymbol{s}})}{(H\widehat{\boldsymbol{s}} - H\boldsymbol{s}^{(p)})^T (HH^T)^{-1} (H\widehat{\boldsymbol{s}} - H\boldsymbol{s}^{(p)})} \tag{5.2b}$$

is the *closed-form* solution of the line search:

$$\alpha_1 = \arg\max_{\alpha} p_{\boldsymbol{y}|\boldsymbol{\theta}}\big(\boldsymbol{y} \,|\, (\widehat{\boldsymbol{s}} + \alpha (\widehat{\boldsymbol{s}} - \boldsymbol{s}^{(p)}), \sigma^2)\big) \tag{5.2c}$$

with the parameter space of $\boldsymbol{\theta}$ extended to Θ_{r_1}, where $r_1 = \dim(\mathrm{supp}(\widehat{\boldsymbol{s}}) \cup \mathrm{supp}(\boldsymbol{s}^{(p)}))$ is the sparsity level of $\widehat{\boldsymbol{s}} + \alpha (\widehat{\boldsymbol{s}} - \boldsymbol{s}^{(p)})$ and σ^2 is an arbitrary positive number, see also (2.4a).

3. Second overrelaxation. Compute the linear combination of $\bar{\boldsymbol{z}}$ and $\boldsymbol{s}^{(p-1)}$:

$$\widetilde{\boldsymbol{z}} = \bar{\boldsymbol{z}} + \alpha_2 (\bar{\boldsymbol{z}} - \boldsymbol{s}^{(p-1)}) \tag{5.3a}$$

where the weight

$$\alpha_2 = \frac{(H\bar{\boldsymbol{z}} - H\boldsymbol{s}^{(p-1)})^T (HH^T)^{-1} (\boldsymbol{y} - H\bar{\boldsymbol{z}})}{(H\bar{\boldsymbol{z}} - H\boldsymbol{s}^{(p-1)})^T (HH^T)^{-1} (H\bar{\boldsymbol{z}} - H\boldsymbol{s}^{(p-1)})} \tag{5.3b}$$

is the closed-form solution of the line search:

$$\alpha_2 = \arg\max_{\alpha} p_{\boldsymbol{y}|\boldsymbol{\theta}}\big(\boldsymbol{y} \,|\, (\bar{\boldsymbol{z}} + \alpha (\bar{\boldsymbol{z}} - \boldsymbol{s}^{(p-1)}), \sigma^2)\big) \tag{5.3c}$$

with the parameter space of $\boldsymbol{\theta}$ extended to Θ_{r_2}, where $r_2 = \dim(\mathrm{supp}(\bar{\boldsymbol{z}}) \cup \mathrm{supp}(\boldsymbol{s}^{(p-1)}))$ is the sparsity level of $\bar{\boldsymbol{z}} + \alpha (\bar{\boldsymbol{z}} - \boldsymbol{s}^{(p-1)})$ and σ^2 is an arbitrary positive number.

4. Thresholding. Threshold $\widetilde{\boldsymbol{z}}$ to the sparsity level r:

$$\widetilde{\boldsymbol{s}} = \mathcal{T}_r(\widetilde{\boldsymbol{z}}) \tag{5.4a}$$

compute the corresponding variance component estimate:

$$\widetilde{\sigma}^2 = (\boldsymbol{y} - H\widetilde{\boldsymbol{s}})^T (HH^T)^{-1} (\boldsymbol{y} - H\widetilde{\boldsymbol{s}})/N \tag{5.4b}$$

and define our final overrelaxation parameter estimate $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{s}}, \widetilde{\sigma}^2)$.

5. Decision (between ECME and thresholded overrelaxation parameter estimates). If $p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\widetilde{\boldsymbol{\theta}}) \geq p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\widehat{\boldsymbol{\theta}})$ or, equivalently, if

$$\widetilde{\sigma}^2 < \widehat{\sigma}^2 \tag{5.5}$$

assign $\boldsymbol{\theta}^{(p+1)} = \widetilde{\boldsymbol{\theta}}$; otherwise, assign $\boldsymbol{\theta}^{(p+1)} = \widehat{\boldsymbol{\theta}}$ and complete *Iteration* $p + 1$.

Iterate until two consecutive sparse-signal estimates $\boldsymbol{s}^{(p)}$ and $\boldsymbol{s}^{(p+1)}$ do not differ significantly. Since $(H\,H^T)^{-1}$ can be pre-computed, our DORE iteration *does not* require matrix inversion; the line searches in the two overrelaxation steps have closed-form solutions and are therefore computationally efficient, see Section V-A for details on computational complexity.

If the rows of the sensing matrix H are orthonormal [i.e. (2.8) holds], Step **1** of the DORE scheme reduces to one IHT step. After Step **1**, we apply two overrelaxations (Steps **2** and **3**) that utilize the sparse signal estimates $\boldsymbol{s}^{(p)}$ and $\boldsymbol{s}^{(p-1)}$ from the two most recent completed DORE iterations. The goal of the overrelaxation steps is to boost the marginal likelihood (2.4a) and accelerate the convergence of the ECME iteration. Using a single overrelaxation step based on the most recent parameter estimate is a common approach for accelerating fixed-point iterations, see [28]. Here, we adopt the idea in [28, Sect. 5.1] and apply the *second overrelaxation*, which mitigates the 'zigzagging' effect caused by the first overrelaxation and thereby converges more rapidly. Our algorithm differs from that in [28, Sect. 5.1], which focuses on continuous parameter spaces with marginal likelihood that is differentiable with respect to the parameters. Unlike [28, Sect. 5.1], here we

- apply overrelaxation steps on parameter spaces with variable dimensions (Steps **2** and **3**),
- threshold the second overrelaxation estimate to ensure that the resulting signal estimate is r-sparse (Step **4**), and
- test the thresholded estimate from Step **4** versus the ECME estimate from Step **1** and adopt the better of the two (Step **5**).

Step **5** ensures that the resulting new parameter estimate $\boldsymbol{\theta}^{(p+1)}$ yields the marginal likelihood function (2.4a) that is higher than or equal to that of the standard ECME step (Step **1**). Therefore, the DORE iteration (5.1)–(5.5) ensures monotonically nondecreasing marginal likelihood between consecutive iteration steps:

$$p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta}^{(p+1)}) \geq p_{\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{y}\,|\,\boldsymbol{\theta}^{(p)}). \tag{5.6}$$

Furthermore, under the conditions of Theorem 1, the DORE iteration converges to a fixed point of the ECME iteration. This convergence result follows from the facts that each DORE iteration contains an ECME step and yields the marginal likelihood (2.4a) that is higher than or equal to that of the standard

ECME step. To see this, consider two consecutive DORE signal estimates $\boldsymbol{s}^{(p)}$ and $\boldsymbol{s}^{(p+1)}$ and the ECME estimate $\widehat{\boldsymbol{s}}$ in *Iteration* $p + 1$. Due to Step **5** of the DORE scheme, we have

$$\mathcal{E}(\boldsymbol{s}^{(p)}) - \mathcal{E}(\boldsymbol{s}^{(p+1)}) \;\geq\; \mathcal{E}(\boldsymbol{s}^{(p)}) - \mathcal{E}(\widehat{\boldsymbol{s}}) \tag{5.7a}$$

$$\geq\; \left[1 - \lambda_{\max}\big(H_A^T\,(H\,H^T)^{-1}\,H_A\big)\right] \|\widehat{\boldsymbol{s}} - \boldsymbol{s}^{(p)}\|_{\ell_2}^2 \tag{5.7b}$$

where $A = \operatorname{supp}(\boldsymbol{s}^{(p)}) \cup \operatorname{supp}(\widehat{\boldsymbol{s}})$. (5.7b) follows from (B.6) of the proof of Theorem 1 in Appendix B. Since the sequence $\mathcal{E}(\boldsymbol{s}^{(p)}) - \mathcal{E}(\boldsymbol{s}^{(p+1)})$ converges to zero and the conditions of Theorem 1 ensure that the term $1 - \lambda_{\max}\big(H_A^T\,(H\,H^T)^{-1}\,H_A\big)$ in (5.7b) is strictly positive (see the proof of Theorem 1 in Appendix B), $\|\widehat{\boldsymbol{s}} - \boldsymbol{s}^{(p)}\|_{\ell_2}^2$ converges to zero as well, implying the convergence of the DORE iteration to an ECME fixed point. In addition, by Theorem 2, the fixed point that DORE converges to is also a local maximum of the concentrated marginal likelihood function (2.6). The near-optimal recovery results (Theorems 3 and 4) in Section IV also apply to DORE and can be easily derived along the lines of the proofs for the ECME algorithm in Appendix D using the facts that each DORE iteration contains an ECME step and yields the marginal likelihood (2.4a) that is higher than or equal to that of the standard ECME step.

DORE Initialization. The parameter estimates $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are obtained by applying two consecutive ECME steps (2.7) to an initial sparse signal estimate $\boldsymbol{s}^{(0)}$.

A. *Computational Complexity and Memory Requirements*

The major computational complexity of the ECME algorithm lies in the matrix-vector multiplications and sorting of $m \times 1$ vectors. Assuming that the common bubble sorting is employed, sorting $\boldsymbol{z}^{(p+1)}$ in (2.7b) requires $\mathcal{O}(m^2)$ operations. There are three matrix-vector multiplications in one ECME iteration, namely $H\boldsymbol{s}^{(p)}$, $(HH^T)^{-1}[H\boldsymbol{s}^{(p)}]$ and $H^T[(HH^T)^{-1}H\boldsymbol{s}^{(p)}]$, which requires $\mathcal{O}(Nm)$, $\mathcal{O}(N^2)$ and $\mathcal{O}(Nm)$ operations, respectively. The intermediate computation results of $(\sigma^2)^{(p+1)}$ in (2.7c) can be stored and used to compute (2.7b) of the next iteration; therefore, this step does not cause additional computation. In summary, the complexity of one ECME iteration is $\mathcal{O}(m^2 + 2Nm + N^2)$. If H has orthonormal rows satisfying (2.8), ECME reduces to the IHT iteration, and in this case $(HH^T)^{-1}[H\boldsymbol{s}^{(p)}]$ is simply $H\boldsymbol{s}^{(p)}$. The computation complexity of one IHT step is therefore $\mathcal{O}(m^2 + 2Nm)$.

For DORE, there are two sorting operations per iteration, one in step **1** and the other in step **4**, requiring $\mathcal{O}(2m^2)$ operations. In one DORE iteration, we need to compute $H^T[(HH^T)^{-1}H\boldsymbol{s}^{(p)}]$, $H\,\widehat{\boldsymbol{s}}$, $(H\,H^T)^{-1}[H\,\widehat{\boldsymbol{s}}]$, $H\,\widetilde{\boldsymbol{s}}$, and $(H\,H^T)^{-1}[H\,\widetilde{\boldsymbol{s}}]$, which require total of $\mathcal{O}(3Nm + 2N^2)$ operations. Note that $H\boldsymbol{s}^{(p-1)}$, $(HH^T)^{-1}[H\boldsymbol{s}^{(p-1)}]$, $H\boldsymbol{s}^{(p)}$, and $(HH^T)^{-1}[H\boldsymbol{s}^{(p)}]$ in (5.1a), (5.2b) and (5.3b) can be adopted from the previous two iterations and do not need to be computed again in the current iteration; in addition, the quantities $H\bar{\boldsymbol{z}}$ and $(HH^T)^{-1}[H\bar{\boldsymbol{z}}]$ in (5.3b) are simple linear combinations of computed terms

and do not require additional matrix-vector computations. To summarize, one DORE iteration requires $\mathcal{O}(2m^2 + 3Nm + 2N^2)$, which is slightly less than twice the complexity of one ECME step. When H has orthonormal rows, we do not need to compute $(H\,H^T)^{-1}\,[H\,\widehat{s}]$ and $(H\,H^T)^{-1}\,[H\,\widetilde{s}]$, which brings the complexity down to $\mathcal{O}(2m^2 + 3Nm)$, slightly less than twice the complexity of one IHT step.

Regarding the memory storage, the largest quantity that ECME (and its special case IHT) and DORE need to store is the sensing matrix H requiring memory storage of order $\mathcal{O}(Nm)$. In large-scale applications, H is typically not explicitly stored but instead appears in the function-handle form [for example, random DFT sensing matrix can be implemented via the fast Fourier transform (FFT)]. In this case, the storage requirement of ECME, IHT and DORE is just $\mathcal{O}(m)$.

Although a single DORE step is about twice more complex than the ECME and IHT steps, it converges in much fewer iterations than the ECME and IHT iterations in the numerical examples in Section VII, see Fig. 3 (b) and Fig. 4 (c).

VI. UNCONSTRAINED SPARSITY SELECTION CRITERION FOR SELECTING r AND THE ADORE ALGORITHM

The ECME and DORE algorithms, as well as most other greedy methods, require the knowledge of sparsity level r as an input. In this section, we propose an sparsity selection criterion and an *automatic double overrelaxation* (ADORE) thresholding algorithm that estimates the signal sparsity from the measurements.

We introduce the following *unconstrained sparsity selection (USS)* objective function for selecting the proper sparsity level r that strikes a balance between the efficiency and accuracy of signal representation:

$$\mathrm{USS}(r) = -\tfrac{1}{2}\,r\,\ln\Big(\frac{N}{m}\Big) - \tfrac{1}{2}\,(N - r - 2)\,\ln\Big(\frac{\widehat{\sigma}^2_{\mathrm{ML}}(r)}{\boldsymbol{y}^T\,(H\,H^T)^{-1}\,\boldsymbol{y}/N}\Big) \tag{6.1}$$

where $\widehat{\sigma}^2_{\mathrm{ML}}(r)$ is the ML estimate of the variance component σ^2 in the parameter space Θ_r, see (2.4b). $\mathrm{USS}(r)$ in (6.1) is developed from the approximate generalized maximum likelihood (GML) objective function in [26, e.q. (13)]; in particular, when $\boldsymbol{y}^T\,(H\,H^T)^{-1}\,\boldsymbol{y}/N = 1$, the two functions are equal up to an additive constant. However, unlike GML, the USS objective function (6.1) is scale-invariant: scaling the measurements $\boldsymbol{y}$ by a nonzero constant does not change $\mathrm{USS}(r)$, which is a desirable property.

Interestingly, the USS objective (6.1) is closely related to the (P$_0$) problem (1.1), as shown by the following theorem.

Theorem 5: Suppose that we have collected a measurement vector $\boldsymbol{y} = H\,\boldsymbol{s}^\diamond$ using a proper sensing matrix H, where $\boldsymbol{s}^\diamond$ is a sparse signal vector having exactly $r^\diamond = \|\boldsymbol{s}^\diamond\|_{\ell_0}$ nonzero elements. If

(1)   the sensing matrix H satisfies the unique representation property (URP) condition (3.5b) and

(2)   the number of measurements N satisfies

$$N \geq \max\{2\,r^\diamond, r^\diamond + 3\} \tag{6.2}$$

then

- USS(r) in (6.1) is *globally and uniquely maximized* at $r = r^\diamond$ and
- the (P_0)-optimal solution and ML sparse signal estimate at $r = r^\diamond$ [i.e. $\widehat{\boldsymbol{s}}_{\mathrm{ML}}(r^\diamond)$, see (2.4b)] are both *unique* and *coincide* with $\boldsymbol{s}^\diamond$.

   Proof: See Appendix E. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

   Theorem 5 shows that the USS objective function *transforms* the constrained optimization problem (P_0) in (1.1) into an equivalent unconstrained problem (6.1) and that USS optimally selects the signal sparsity level r that allows accurate signal representation with as few nonzero signal elements as possible.

   In the practical scenarios where $r^\diamond \geq 3$, condition (2) of Theorem 5 reduces to $N \geq 2\,r^\diamond$, which is the condition required to ensure the uniqueness of the (P_0) problem, see [8, Theorem 2].

   In the following, we use DORE to approximately evaluate the USS objective function and apply this approximate USS criterion to automatically select the signal sparsity level.

A. The ADORE Algorithm for Unknown Sparsity Level r

We approximate the USS objective function (6.1) by replacing the computationally intractable ML estimate $\widehat{\sigma}_{\mathrm{ML}}^2(r)$ with its DORE estimate. Maximizing this approximate USS objective function with respect to r by an exhaustive search may be computationally expensive because we need to apply a full DORE iteration for each sparsity level r in the set of integers between 0 and $N/2$.[2] Here, we propose the ADORE algorithm that applies the *golden-section search* [34, Sec. 4.5.2.1] to maximize the approximate USS objective function with respect to r, with the initial search boundaries set to 0 and $\lceil N/2 \rceil$. Note that USS$(0) = 0$ assuming that $\boldsymbol{y} \neq \boldsymbol{0}_{N \times 1}$, which is of practical interest. For each candidate $0 < r \leq \lceil N/2 \rceil$, we estimate $\widehat{\sigma}_{\mathrm{ML}}^2(r)$ using the DORE iteration. After running one golden sectioning step, the length of the new search interval is approximately 0.618 of the previous interval (rounded to the closest integer). The search process ceases when the desired resolution L is reached, i.e. when the searching interval becomes shorter than the prescribed resolution level L. Therefore, ADORE requires roughly $1.4\left[\log_2(N/L) - 1\right]$ full DORE iterations. For the golden-section search to find the exact maximum of (6.1), USS(r) must be *unimodal* in r, which is not true in general. Hence, ADORE maximizes (6.1) only approximately, yielding r_{ADORE}; then, our ADORE sparse-signal estimate is equal to the corresponding DORE estimate at $r = r_{\mathrm{ADORE}}$.

VII. NUMERICAL EXAMPLES

We now compare our proposed methods in Sections V and VI with existing large-scale sparse reconstruction techniques using two image recovery experiments, with purely and approximately sparse signals,

---

[2]Note that $N/2$ is the largest value of the sparsity level r for which reasonable reconstruction is possible from N measurements; otherwise, the (P_0) and ML estimates of the sparse signal may not be unique, see e.g. [8, Theorem 2].

respectively. In particular, we compare

- the DORE and ADORE schemes initialized by the zero sparse signal estimate:

$$s^{(0)} = \mathbf{0}_{m \times 1} \tag{7.1}$$

  with ADORE search resolution set to $L = 500$ and MATLAB implementations available at http://home.eng.iastate.

- the IHT and NIHT schemes in [20] and [21], initialized by the zero sparse signal estimate $s^{(0)}$ in (7.1);

- the automatic hard thresholding (AHT) method in [26] using the moving-average window length 100, initialized with $z_{\mathrm{init}} = \mathbf{0}_{m \times 1}$ and $r_{\mathrm{init}} = 1$;

- the debiased gradient-projection for sparse reconstruction method in [13, Sec. III.B] with the convergence threshold $\mathtt{tolP} = 10^{-5}$ and regularization parameter set to

  (i) $\tau = 0.1 \, \|H^T y\|_{\ell_\infty}$, suggested in [13, e.q. (22)] (labeled GPSR$_0$) and

  (ii) $\tau = 0.001 \, \|H^T y\|_{\ell_\infty}$, obtained by manual tuning for good performance in the following two numerical examples (labeled GPSR);

- the minimum-norm signal estimate (labeled MN):

$$\widehat{s}_{\mathrm{MN}} = H^T \left(H \, H^T \right)^{-1} y \tag{7.2}$$

  which achieves zero squared residual error by ignoring sparsity.

For the DORE, ADORE, IHT and NIHT iterations, we use the following convergence criterion[3]:

$$\|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2 \, / \, m < 10^{-14}. \tag{7.3}$$

The sensing matrix H has the following structure (see e.g. [5, eq. (2) and Fig. 1]):

$$H = \Phi \, \Psi \tag{7.4}$$

where Φ is an $N \times m$ *sampling matrix* and Ψ is an appropriate $m \times m$ orthogonal *sparsifying transform matrix*. In our examples presented here, Ψ are inverse discrete wavelet transform (DWT) matrices [41].

For an underlying image $\Psi \, s$, the signal vector s is the wavelet coefficient vector of the image. Our performance metric is the peak signal-to-noise ratio (PSNR) of a reconstructed image $\Psi \, \widehat{s}$, where $\widehat{s}$ is the estimated wavelet coefficients vector:

$$\mathrm{PSNR \ (dB)} = 10 \, \log_{10} \left\{ \frac{[(\Psi \, s)_{\mathrm{MAX}} - (\Psi \, s)_{\mathrm{MIN}}]^2}{\| \Psi \, \widehat{s} - \Psi \, s \|_{\ell_2}^2 / m} \right\} = 10 \, \log_{10} \left\{ \frac{[(\Psi \, s)_{\mathrm{MAX}} - (\Psi \, s)_{\mathrm{MIN}}]^2}{\| \widehat{s} - s \|_{\ell_2}^2 / m} \right\} \tag{7.5}$$

where $(\Psi \, s)_{\mathrm{MIN}}$ and $(\Psi \, s)_{\mathrm{MAX}}$ denote the smallest and largest elements of $\Psi \, s$.

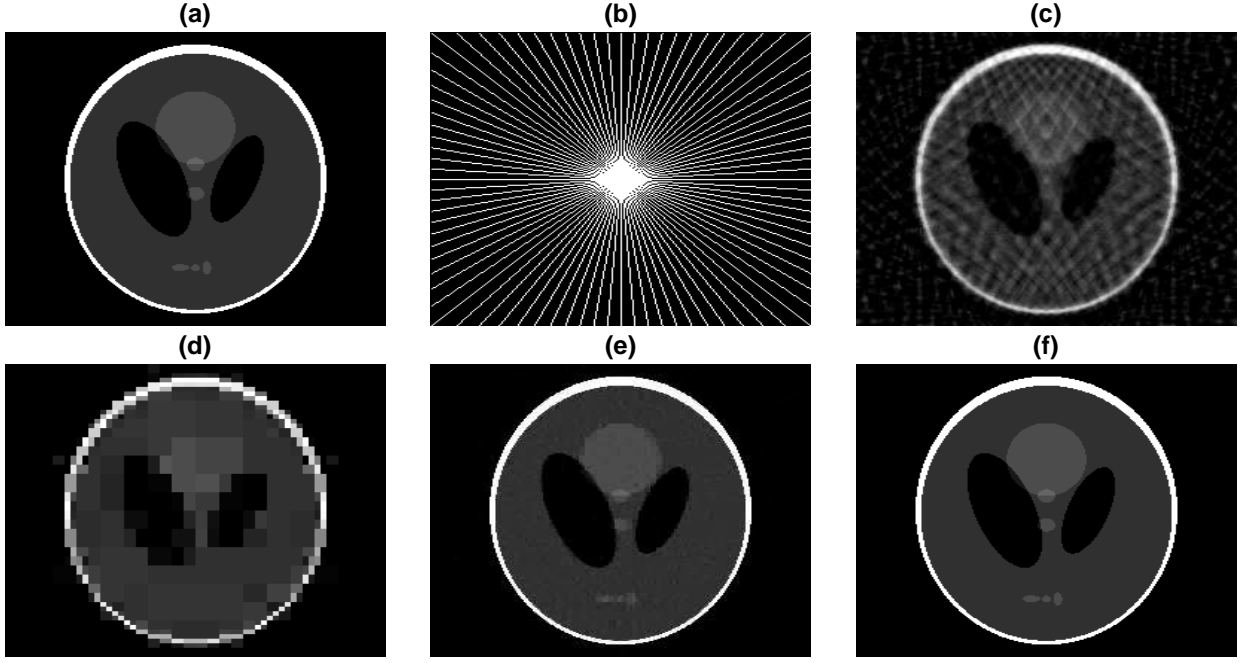[3]To implement the IHT and NIHT schemes, we incorporated the convergence criterion (7.3) into the corresponding MATLAB codes from the `sparsify` toolbox at http://www.see.ed.ac.uk/~tblumens/sparsify/sparsify.html.

Fig. 2.   (a) The size-256^2 Shepp-Logan phantom, (b) a star-shaped sampling domain in the frequency plane containing 44 radial lines, and (c) the filtered back-projection (minimum-norm) reconstruction, (d) GPSR$_0$ reconstruction, (e) GPSR reconstruction, and (f) the almost perfect reconstruction achieved by all hard-thresholding schemes, for the sampling pattern in (b).

A. Tomographic Image Reconstruction

Consider the reconstruction of the Shepp-Logan phantom of size $m = 256^2$ in Fig. 2 (a) from tomographic projections. The elements of $\boldsymbol{y}$ are 2-D discrete Fourier transform (DFT) coefficients of the phantom sampled over a star-shaped domain, as illustrated in Fig. 2 (b); see also [2], [21], and [26]. Therefore, the sampling matrix Φ is constructed using selected rows of the DFT matrix that yield the corresponding DFT coefficients of the phantom image within the star-shaped domain. In this example, we select the inverse Haar (Daubechies-2) DWT matrix to be the orthogonal sparsifying transform matrix Ψ. The Haar wavelet transform coefficients of the phantom image in Fig. 2 (a) are sparse, with $\|\boldsymbol{s}\|_{\ell_0} = 3769 \approx 0.06\, m$, where the true signal vector $\boldsymbol{s}$ consists of the Haar wavelet transform coefficients of the phantom in Fig. 2 (a).

For our choices of Φ and Ψ, the rows of H are orthonormal, i.e. (2.8) holds, implying that IHT is equivalent to the ECME iteration in Section II-A. DORE, IHT, and NIHT require knowledge of the signal sparsity level r; in this example, we set r to the true signal support size:

$$r = 3769. \tag{7.6}$$

In contrast, the ADORE and AHT methods are *automatic* and *estimate* r from the measurements using the USS and GML selection criteria, respectively.

Figs. 2 (c)–(f) show the images reconstructed by the above methods using the 44 radial-line sampling pattern in Fig. 2 (b), which corresponds to $N/m = 0.163$. In this example, the MN signal estimate (7.2)
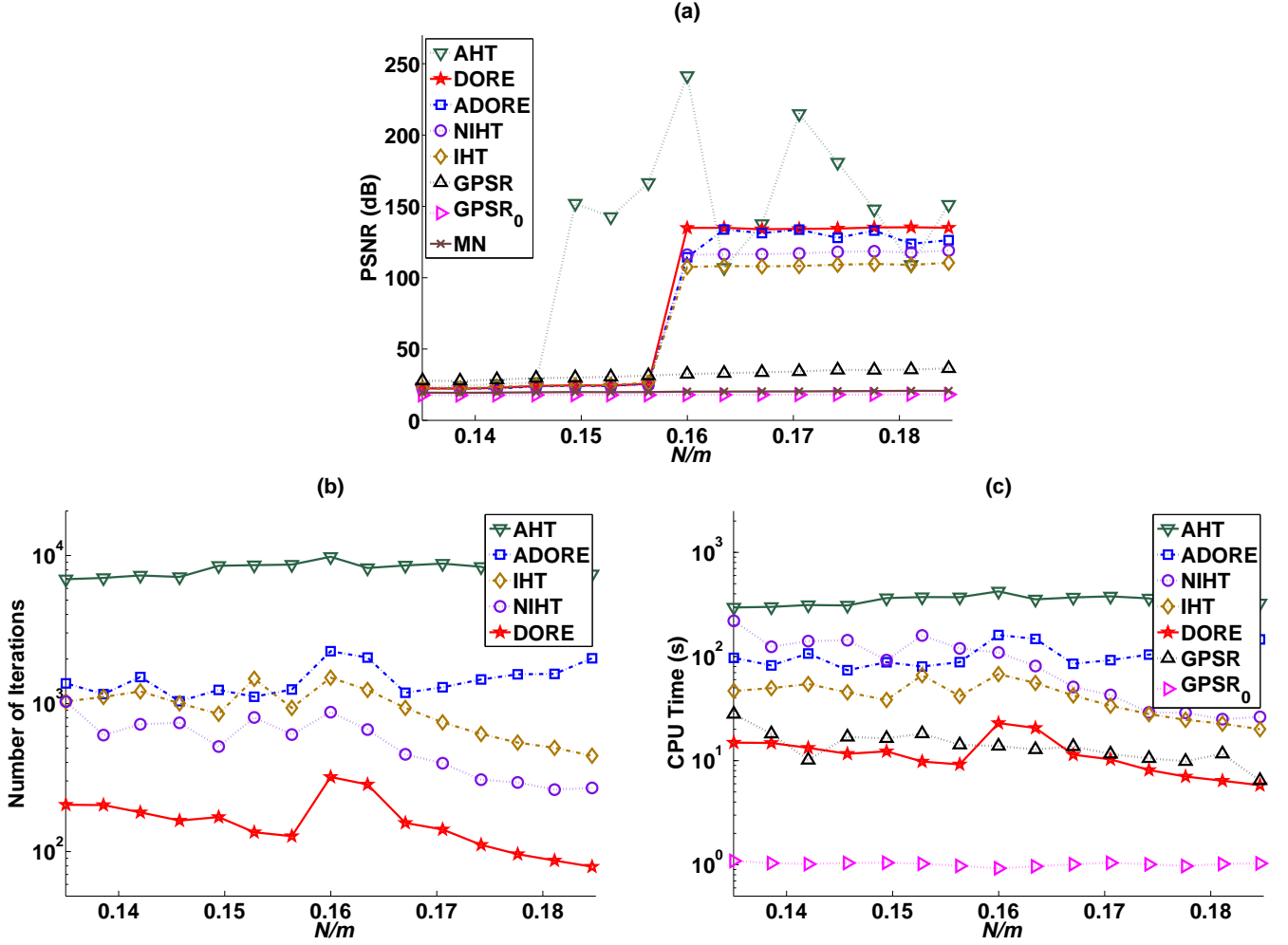
Fig. 3. (a) PSNR, (b) number of iterations, and (c) CPU time as the functions of the normalized number of measurements N/m for phantom image reconstruction.

is also the *filtered back-projection estimate* obtained by setting the unobserved DFT coefficients to zero and taking the inverse DFT, see [2]. Here, all hard-thresholding methods (DORE, IHT, NIHT, ADORE, and AHT) achieve almost perfect reconstructions of the original phantom image with PSNRs over 100 dB, in contrast with the MN (filtered back-projection) and GPSR methods that achieve inferior reconstructions with PSNRs 20.2 dB for the MN, 33.0 dB for GPSR, and 17.9 dB for GPSR$_0$ estimates.

Fig. 3 shows (a) the PSNRs, (b) numbers of iterations, and (c) CPU times of the above methods as we change N/m by varying the number of radial lines in our star-shaped partial Fourier sampling pattern. In this example, all hard-thresholding methods have significantly sharper phase transitions than the manually tuned GPSR, and outperform GPSR after the phase transitions. AHT exhibits the phase transition at $N/m \approx 0.15$; the phase transitions of the other hard thresholding methods occur at $N/m \approx 0.16$. ADORE performs as well as the DORE, IHT, and NIHT methods that require prior knowledge of the signal sparsity level. Indeed, the USS criterion accurately selects the signal sparsity level in this case, which is consistent with

the essence of Theorem 5. Among all hard-thresholding methods, DORE needs the smallest number of iterations to converge and is also the fastest in terms of the CPU time. DORE needs 4.4 to 10.9 times less iterations than IHT and 2.3 to 6 times less iterations than NIHT; in terms of the CPU time, DORE is 2.7 to 6.7 times faster than IHT and 3.6 to 16.3 times faster than NIHT. The CPU times of DORE, IHT, and ADORE *per iteration* are approximately constant as N/m varies. One DORE step is about twice slower than one IHT step, validating the computational complexity analysis in Section V-A.

We now compare the two automatic thresholding methods (AHT and ADORE) in this example: ADORE requires 3.7 to 7.7 times less iterations and is 2.2 to 4.6 times faster than AHT. We note that AHT's computational complexity does not scale well with the increase of the signal size and its support, which is the case considered in the following example, where we increase the signal size four times, to $m = 512^2$.

B. Lena Reconstruction From Compressive Samples

We now reconstruct the standard Lena image of size $m = 512^2$ in Fig. 4 (a) from compressive samples. In this example, we select the *structurally random sampling matrices* Φ proposed in [42] and the inverse Daubechies-6 DWT matrix to be the orthogonal sparsifying transform matrix Ψ. The wavelet coefficients of the Lena image are only approximately sparse. If we have a parameter estimate $\widehat{\boldsymbol{\theta}}(r) = \big(\widehat{\boldsymbol{s}}(r), \widehat{\sigma}^2(r)\big)$, we can construct the following empirical Bayesian estimate of the missing data vector $\boldsymbol{z}$:

$$\mathrm{E}_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{z}\,|\,\boldsymbol{y}, \widehat{\boldsymbol{\theta}}(r)] = \widehat{\boldsymbol{s}}(r) + H^T(HH^T)^{-1}[\boldsymbol{y} - H\,\widehat{\boldsymbol{s}}(r)]. \tag{7.7}$$

Unlike $\widehat{\boldsymbol{s}}(r)$, the empirical Bayesian estimate (7.7) is not r-sparse in general, and is therefore preferable for reconstructing approximately sparse signals that have many small-magnitude signal coefficients.

For our choices of Φ and Ψ, the rows of H are orthonormal, i.e. (2.8) holds, and IHT is equivalent to the ECME iteration in Section II-A. For all hard thresholding methods, we apply the empirical Bayesian estimate (7.7), with $\widehat{\boldsymbol{\theta}}(r)$ equal to the parameter estimates obtained upon their convergence. We chose the sparsity level

$$r = 10000 \approx 0.038\,m \tag{7.8}$$

to implement the DORE, IHT, and NIHT iterations.

Figs. 4 (b)–(d) show the PSNRs, numbers of iterations, and CPU times of various methods as functions of the normalized number of measurements (subsampling factor) N/m. Here, we do not include the AHT and MN estimates in the simulation results because, in this example, AHT is very slow compared with the other approaches and does not outperform them in terms of reconstruction performance, and the MN estimates (7.2) are poor; indeed, the PSNRs of the minimum-norm estimates vary between 14.21 dB and 16.24 dB for the range of N/m in Figs. 4 (b)–(d).
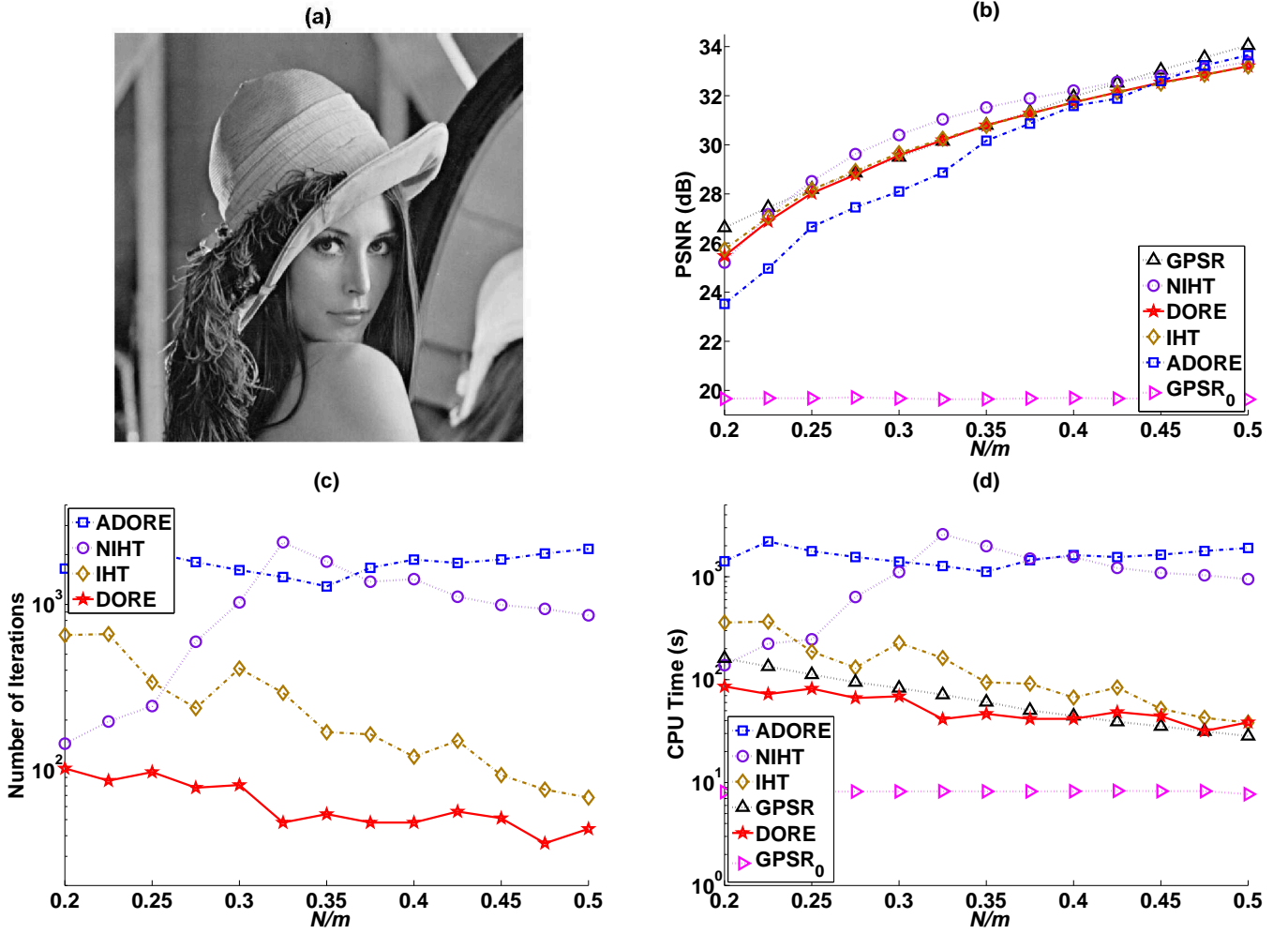
Fig. 4.   (a) The 512×512 Lena image, and (b) PSNR, (c) number of iterations, and (d) CPU time as the functions of the normalized number of measurements N/m for Lena image reconstruction.

Unlike the phantom reconstruction example in Section VII-A, here the underlying signal (the vector of the wavelet coefficients of the Lena image) is *not* strictly sparse and, consequently,

- the difference in reconstruction accuracy between the *manually tuned* convex GPSR method and hard thresholding methods is significantly smaller: compare Fig. 4 (b) with Fig. 3 (a);

- the achieved PSNRs of all methods are significantly smaller as well, even though the subsampling factor N/m ranges over a fairly wide interval, between 0.2 and 0.5.

The importance of tuning the GPSR's regularization parameter τ is evident from Fig. 4 (b): GPSR$_0$ reconstructs the signal poorly compared with the tuned GPSR. We point out that it is *not known* how to manually tune τ in practical cases where the ground-truth image in Fig. 4 (a) is not available. In contrast, our ADORE algorithm automatically selects the sparsity level and performs similarly to the other methods that require careful tuning; ADORE is particularly competitive when the number of measurements is fairly

large, see Fig. 4 (b). Therefore, the USS model selection criterion is quite effective in this practical example where the underlying signal is not strictly sparse. Figs. 4 (c) and (d) show that DORE requires the smallest number of iterations and CPU time among the hard thresholding methods, and that it is also faster than the manually tuned GPSR method for $N/m < 0.4$. When $N/m > 0.3$, the CPU time of the *automatic* ADORE method (which employs multiple DORE iterations) is comparable to that of the *tuned* NIHT method.

VIII. Concluding Remarks

We proposed a probabilistic framework for sparse signal reconstruction in underdetermined linear models where the regression coefficient vector consists of a sparse deterministic component and a random Gaussian component. We developed three hard thresholding methods based on this framework: ECME, DORE, and ADORE. We showed that, under certain mild conditions, ECME converges to a local maximum of the concentrated marginal likelihood for the above probabilistic model. Our ECME convergence conditions are invariant to invertible linear transforms of either the rows or the columns of the sensing matrix. To develop our near-optimal recovery results for the ECME and DORE methods, we introduced new measures of the sensing matrix's reconstruction ability: sparse subspace quotient (SSQ) and minimum SSQ. Minimum SSQ is more flexible than the well-established restricted isometry property (RIP) and coherence measures: it is invariant to invertible linear transforms of the rows of the sensing matrix. When the minimum $2r$-SSQ is sufficiently large, our ECME for sparsity level r *perfectly* recovers the true r-sparse signal from the noiseless measurements and estimates the best r-term approximation of an arbitrary non-sparse signal from noisy measurements within a bounded error. The DORE algorithm interleaves two overrelaxation steps with one ECME step and significantly accelerates the convergence of the ECME iteration. To automatically estimate the sparsity level from the data, we proposed the unconstrained sparsity selection (USS) criterion and utilized it to develop the automatic ADORE scheme that does not require prior knowledge of the signal sparsity level.

Since only a single choice (7.1) is used to initialize DORE and ADORE, their PSNR curves in Section VII are only lower bounds on the PSNRs achievable by these methods. The reconstruction performances of these methods can be improved by using multiple initial values, where the improvement is particularly significant for purely sparse signals: our preliminary results indicate that, in terms of reconstruction accuracy, DORE with multiple initial values outperforms AHT in the phantom example [see Fig. 3 (a)] and can slightly outperform the manually tuned GPSR in the Lena example [see Fig. 4 (b)]. Full details of the multiple initialization scheme and its reconstruction performance will be published elsewhere. Further research will also include:

- analyzing the convergence speed of the DORE algorithm,

- looking for systematic means of generating sensing matrices that have large minimum SSQ, and

- applying our probabilistic framework to develop sparse signal reconstruction methods for quantized measurements.

APPENDIX

APPENDIX A
ECME ALGORITHM DERIVATION

Consider the following hierarchical two-stage model:

$$p_{\boldsymbol{y}\,|\,\boldsymbol{z}}(\boldsymbol{y}\,|\,\boldsymbol{z}) \;=\; \mathcal{N}(\boldsymbol{y}\,;\, H\,\boldsymbol{z}, C) \tag{A.1a}$$

$$p_{\boldsymbol{z}\,|\,\boldsymbol{\theta}}(\boldsymbol{z}\,|\,\boldsymbol{\theta}) \;=\; \mathcal{N}(\boldsymbol{z}\,;\, \boldsymbol{s}, \sigma^2\, I_m) \tag{A.1b}$$

where $\boldsymbol{z}$ is the vector of missing data and C is a known noise covariance matrix. For $C = 0_{N\times N}$, this model reduces to that in (2.1a)–(2.1b) in Section II.

We will first derive an EM step for estimating $\boldsymbol{s}$ under the above general model and then set $C = 0_{N\times N}$ to reduce it to the EM step in Section II. The *complete-data likelihood function* of the measurements $\boldsymbol{y}$ and the missing data $\boldsymbol{z}$ given $\boldsymbol{\theta} = (\boldsymbol{s}, \sigma^2) \in \Theta_r$ follows from (A.1a) and (A.1b):

$$p_{\boldsymbol{z},\boldsymbol{y}\,|\,\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{y}\,|\,\boldsymbol{\theta}) \;=\; \frac{\exp[-\frac{1}{2}\,(\boldsymbol{y} - H\,\boldsymbol{z})^T\,C^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{z})]}{\sqrt{\det(2\,\pi\,C)}} \cdot \frac{\exp(-\frac{1}{2}\,\|\boldsymbol{z} - \boldsymbol{s}\|_{\ell_2}^2/\sigma^2)}{\sqrt{(2\,\pi\,\sigma^2)^m}}. \tag{A.2}$$

From (A.2), the conditional pdf of $\boldsymbol{z}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$ is

$$p_{\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}) = \mathcal{N}\Big(\boldsymbol{z}\,;\, \boldsymbol{s} + \sigma^2\,H^T\,(C + \sigma^2\,H\,H^T)^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s}), \sigma^2\,I_m - (\sigma^2)^2\,H^T\,(C + \sigma^2\,H\,H^T)^{-1}\,H\Big) \tag{A.3}$$

see [33, Theorem 11.1]. Assume that the parameter estimate $\boldsymbol{\theta}^{(p)} = \big(\boldsymbol{s}^{(p)}, (\sigma^2)^{(p)}\big)$ is available; then, in *Iteration $p + 1$*, the E and M steps for estimating $\boldsymbol{s}$ simplify to

$$\boldsymbol{z}^{(p+1)} = \mathrm{E}_{\,\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{z}\,|\,\boldsymbol{y}, \boldsymbol{\theta}^{(p)}] = \boldsymbol{s}^{(p)} + (\sigma^2)^{(p)}\,H^T\,\big[C + (\sigma^2)^{(p)}\,H\,H^T\big]^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s}^{(p)}) \tag{A.4a}$$

and

$$\boldsymbol{s}^{(p+1)} = \arg\min_{\boldsymbol{s}\in\mathcal{S}_r} \|\boldsymbol{z}^{(p+1)} - \boldsymbol{s}\|_{\ell_2}^2 = \mathcal{T}_r(\boldsymbol{z}^{(p+1)}). \tag{A.4b}$$

Setting $C = 0_{N\times N}$ in (A.4a) and (A.4b) yields (2.7a) and (2.7b), which are not dependent on $(\sigma^2)^{(p)}$.

APPENDIX B
PROOF OF THEOREM 1

We first prove Lemma 5, which will be used in the proof of Theorem 1.

Lemma 5: Assume that the sensing matrix H satisfies the URP condition, see also (3.5b). For an index set $A \subset \{1, 2, \ldots, m\}$,

(a)   if

$$0 < \dim(A) \leq N \tag{B.1a}$$

then

$$\lambda_{\min}\left(H_A^T \left(H \, H^T \right)^{-1} H_A \right) > 0, \tag{B.1b}$$

(b)   if

$$0 < \dim(A) \leq m - N \tag{B.2a}$$

then

$$\lambda_{\max}\left(H_A^T \left(H \, H^T \right)^{-1} H_A \right) < 1. \tag{B.2b}$$

Proof: The conditions (3.5b) and (B.1a) imply that all columns of H_A are linearly independent; therefore, $H_A^T (HH^T)^{-1} H_A$ is a full-rank positive definite matrix, and (B.1b) follows.

We now assume (B.2a) and show (B.2b). Observe that

$$\lambda_{\max}\left(H_A^T \left(H \, H^T \right)^{-1} H_A \right) = \lambda_{\max}\left((H \, H^T)^{-1} H_A H_A^T \right) = \lambda_{\max}\left(I_N - (H \, H^T)^{-1} H_{A^c} H_{A^c}^T \right) \tag{B.3}$$

where

$$A^c \triangleq \{1, 2, \ldots, m\} \setminus A \tag{B.4}$$

defines the index index set complementary to A. Since $\dim(A^c) = m - \dim(A) \geq N$, $(H \, H^T)^{-1} H_{A^c} H_{A^c}^T$ is positive definite; therefore,

$$\lambda_{\max}\left(H_A^T \left(H \, H^T \right)^{-1} H_A \right) = 1 - \lambda_{\min}\left((H \, H^T)^{-1} H_{A^c} H_{A^c}^T \right) < 1 \tag{B.5}$$

and (B.2b) follows. $\square$

Now, we prove Theorem 1.

Proof of Theorem 1: We now prove that our ECME iteration converges to its fixed point. If $s^{(p+1)} = s^{(p)}$, the convergence to a fixed point immediately follows. Therefore, without loss of generality, we assume $s^{(p+1)} \neq s^{(p)}$. Since $\mathcal{E}(s^{(p)})$ in (3.1) converges to a limit, $\mathcal{E}(s^{(p)}) - \mathcal{E}(s^{(p+1)})$ converges to zero. Now,

$$
\begin{align}
\mathcal{E}(s^{(p)}) - \mathcal{E}(s^{(p+1)}) &= \mathcal{Q}(s^{(p)} \,|\, s^{(p)}) - \mathcal{H}(s^{(p)} \,|\, s^{(p)}) - [\, \mathcal{Q}(s^{(p+1)} \,|\, s^{(p)}) - \mathcal{H}(s^{(p+1)} \,|\, s^{(p)}) \,] \tag{B.6a} \\
&\geq (s^{(p+1)} - s^{(p)})^T \,[I_m - H^T\,(HH^T)^{-1}\,H]\,(s^{(p+1)} - s^{(p)}) \tag{B.6b} \\
&= (s_A^{(p+1)} - s_A^{(p)})^T \,[I_{\dim(A)} - H_A^T\,(HH^T)^{-1}\,H_A]\,(s_A^{(p+1)} - s_A^{(p)}) \tag{B.6c} \\
&= \left[1 - \frac{(s_A^{(p+1)} - s_A^{(p)})^T\,H_A^T\,(HH^T)^{-1}\,H_A\,(s_A^{(p+1)} - s_A^{(p)})}{\|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2}\right] \|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2 \tag{B.6d} \\
&\geq \left[1 - \lambda_{\max}\big(H_A^T(HH^T)^{-1}H_A\big)\right] \|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2 \tag{B.6e}
\end{align}
$$

where $A = \operatorname{supp}(s^{(p)}) \cup \operatorname{supp}(s^{(p+1)})$. Here, (B.6a) follows from (3.2a), (B.6b) follows by (3.3a) and the fact that $\mathcal{H}(s^{(p)} \,|\, s^{(p)}) = 0$, (B.6c) is obtained by using the identities $\|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2 = \|s_A^{(p+1)} - s_A^{(p)}\|_{\ell_2}^2$ and $H\,(s^{(p+1)} - s^{(p)}) = H_A\,(s_A^{(p+1)} - s_A^{(p)})$, and (B.6e) follows by using the Rayleigh-quotient property [35, Theorem 21.5.6]. Note that $0 < \dim(A) \leq 2\,r \leq m - N$, where the second inequality follows from (3.5a). Therefore, (B.2a) holds and (B.2b) in Lemma 5 implies that the term $1 - \lambda_{\max}\big(H_A^T(HH^T)^{-1}H_A\big)$ in (B.6e) is strictly positive. Since $\mathcal{E}(s^{(p)}) - \mathcal{E}(s^{(p+1)})$ converges to zero, then $\|s^{(p+1)} - s^{(p)}\|_{\ell_2}^2$ converges to zero as well. Finally, the claim of monotonicity of convergence follows from the discussion in Section III prior to Theorem 1. This completes the proof. $\square$

APPENDIX C
PROOFS OF LEMMA 1, LEMMA 2 AND THEOREM 2

Proof of Lemma 1: The proof is by contradiction. Suppose that there exists an index $i \in \{1, 2, \ldots, m\}$ satisfying (3.6a), but not (3.6b); without loss of generality, assume that the following partial derivative is positive:

$$
\frac{\partial f(s^\star)}{\partial s_i} \;\triangleq\; \left.\frac{\partial f(s)}{\partial s_i}\right|_{s = s^\star} = \lim_{\epsilon \to 0} \frac{f(s^\star + \epsilon\,e_i) - f(s^\star)}{\epsilon} > 0 \tag{C.1}
$$

where e_i is the ith column of I_m. By the definition of the limit, there exists a $\delta > 0$ such that, for all $\epsilon \in (0, \delta)$,

$$
\left| \frac{f(s^\star + \epsilon\,e_i) - f(s^\star)}{\epsilon} - \frac{\partial f(s^\star)}{\partial s_i} \right| < \tfrac{1}{2}\frac{\partial f(s^\star)}{\partial s_i} \tag{C.2a}
$$

$$
\left| \frac{f(s^\star - \epsilon\,e_i) - f(s^\star)}{-\epsilon} - \frac{\partial f(s^\star)}{\partial s_i} \right| < \tfrac{1}{2}\frac{\partial f(s^\star)}{\partial s_i} \tag{C.2b}
$$

and, therefore,

$$
f(s^\star + \epsilon\,e_i) > f(s^\star) + \tfrac{1}{2}\epsilon\,\frac{\partial f(s^\star)}{\partial s_i} > f(s^\star) \tag{C.3a}
$$

$$
f(s^\star - \epsilon\,e_i) < f(s^\star) - \tfrac{1}{2}\epsilon\,\frac{\partial f(s^\star)}{\partial s_i} < f(s^\star). \tag{C.3b}
$$

For all $\epsilon \in (0, \delta)$, the vectors $\boldsymbol{s}^\star + \epsilon \, \boldsymbol{e}_i$ and $\boldsymbol{s}^\star - \epsilon \, \boldsymbol{e}_i$ are r-sparse, $f(\boldsymbol{s}^\star + \epsilon \, \boldsymbol{e}_i)$ is larger than $f(\boldsymbol{s}^\star)$, and $f(\boldsymbol{s}^\star - \epsilon \, \boldsymbol{e}_i)$ is smaller than $f(\boldsymbol{s}^\star)$, which contradicts the assumption that $\boldsymbol{s}^\star$ is an r-local maximum or minimum point. $\qquad \square$

Before proving Lemma 2, we prove the following useful result that will be used in the proof of Lemma 2.

Lemma 6: For any r-sparse vector $\boldsymbol{s}' \in \mathcal{S}_r$, there exists a $\delta > 0$ such that, for all $\boldsymbol{s} \in \mathcal{S}_r$ satisfying $\|\boldsymbol{s} - \boldsymbol{s}'\|_{\ell_2} < \delta$, we have

$$\dim\big(\text{supp}(\boldsymbol{s}) \cup \text{supp}(\boldsymbol{s}')\big) \leq r. \tag{C.4}$$

Proof: The proof is by contradiction. First, define $A = \text{supp}(\boldsymbol{s})$ and $A' = \text{supp}(\boldsymbol{s}')$. Suppose that, for all $\delta > 0$, there exists a $\boldsymbol{s} \in \mathcal{S}_r$ satisfying

$$\|\boldsymbol{s} - \boldsymbol{s}'\|_{\ell_2} < \delta \tag{C.5a}$$

and

$$\dim(A \cup A') > r. \tag{C.5b}$$

Since $\dim(A) \leq r$,

$$\dim\big(A' \cap A^c\big) = \dim\big(A \cup A'\big) - \dim\big(A\big) > r - r = 0 \tag{C.6}$$

implying that the set $A' \cap A^c$ is not empty, see also the definition of the complementary index set in (B.4). Choose δ to be half the magnitude of the smallest nonzero element in $\boldsymbol{s}'$:

$$\delta = \tfrac{1}{2} \min_{i \in A'} |s_i'| \tag{C.7}$$

Now,

$$\|\boldsymbol{s} - \boldsymbol{s}'\|_{\ell_2} \geq \big\|\boldsymbol{s}'_{A' \cap A^c}\big\|_{\ell_2} \geq \min_{i \in A'} |s_i'| > \delta \tag{C.8}$$

which contradicts (C.5a). Therefore, for a positive number δ in (C.7), no r-sparse vector $\boldsymbol{s}$ can satisfy the conditions (C.5a) and (C.5b) simultaneously. $\qquad \square$

Lemma 6 shows that, in a sufficiently small neighborhood of an r-sparse vector $\boldsymbol{s}'$, the support sets of all other r-sparse vectors $\boldsymbol{s}$ significantly overlap with the support set of $\boldsymbol{s}'$. In particular, if $\boldsymbol{s}'$ has exactly r nonzero elements (i.e. $\|\boldsymbol{s}'\|_{\ell_0} = r$), then all other r-sparse vectors in its sufficiently small neighborhood must have the same support set as $\boldsymbol{s}'$. If $\boldsymbol{s}'$ has less than r nonzero elements, an r-sparse vector $\boldsymbol{s}$ in this neighborhood can contain a few inconsistent elements that do not belong to the support set of $\boldsymbol{s}'$ *as long as* (C.4) is satisfied. We now prove Lemma 2.

Proof of Lemma 2: We first consider the case of an r-local maximum of $f(s)$ and assume that conditions (1) and (2) hold for a point $s^\star \in \mathcal{S}_r$. By condition (2), for the positive number δ_1, the Hessian matrix is negative semidefinite around $s^\star$ for all $s \in \mathcal{S}_r$ satisfying $\|s - s^\star\|_{\ell_2} < \delta_1$. By Lemma 6, for any r-sparse vector $s^\star$, there exists a $\delta_2 > 0$ such that, for all $s \in \mathcal{S}_r$ satisfying $\|s - s^\star\|_{\ell_2} < \delta_2$, we have

$$\dim(\operatorname{supp}(s) \cup \operatorname{supp}(s^\star)) \leq r. \tag{C.9}$$

Now, for $\delta = \min\{\delta_1, \delta_2\}$, consider any $s \in \mathcal{S}_r$ satisfying $\|s - s^\star\|_{\ell_2} < \delta$, and expand $f(s)$ around $s^\star$ using the Taylor series with Lagrange's form of the remainder [36, p. 243]:

$$
\begin{align}
f(s) - f(s^\star) &= (s - s^\star)^T \frac{\partial f(s)}{\partial s}\bigg|_{s=s^\star} + \tfrac{1}{2}(s - s^\star)^T \frac{\partial^2 f(s)}{\partial s \partial s^T}\bigg|_{s=s^\star + c(s-s^\star)} (s - s^\star) \tag{C.10a} \\
&\leq (s - s^\star)^T \frac{\partial f(s)}{\partial s}\bigg|_{s=s^\star} \tag{C.10b} \\
&= \sum_{i \in \operatorname{supp}(s) \cup \operatorname{supp}(s^\star)} (s_i - s_i^\star) \frac{\partial f(s)}{\partial s_i}\bigg|_{s=s^\star} \tag{C.10c} \\
&= 0 \tag{C.10d}
\end{align}
$$

where $c \in (0,1)$. Since the vector $s^\star + c(s - s^\star)$ is r-sparse and satisfies $\|s^\star + c(s - s^\star) - s^\star\|_{\ell_2} < \delta$, the Hessian in (C.10a) is negative-semidefinite and (C.10b) follows. Condition (1) of Lemma 2 and (C.9) imply that the partial derivatives in (3.7) are zero for all coordinates with indices $i \in \operatorname{supp}(s) \cup \operatorname{supp}(s^\star)$, and (C.10d) follows. Now, we have a $\delta = \min\{\delta_1, \delta_2\} > 0$ such that, for all $s \in \mathcal{S}_r$ satisfying $\|s - s^\star\|_{\ell_2} < \delta$, $f(s) \leq f(s^\star)$; therefore $s^\star$ is an r-local maximum.

If the Hessian matrix $\frac{\partial^2 f(s)}{\partial s \partial s^T}$ is positive semidefinite around $s^\star$, then $\frac{\partial^2 [-f(s)]}{\partial s \partial s^T}$ is negative semidefinite around $s^\star$. Therefore, $s^\star$ is an r-local maximum of $-f(s)$, and, by Definition 1, $s^\star$ is an r-local minimum of $f(s)$. $\qquad\square$

We are now ready to show Theorem 2.

Proof of Theorem 2: Since $\theta^\star = (s^\star, (\sigma^2)^\star)$ is a fixed point of the ECME iteration, we have

$$s^\star = \arg\min_{s \in \mathcal{S}_r} \mathcal{Q}(s \mid s^\star) = \arg\min_{s \in \mathcal{S}_r} \|s - [s^\star + H^T(HH^T)^{-1}(y - Hs^\star)]\|_{\ell_2}^2 \tag{C.11}$$

see (2.7b) and (3.2b).

We first show that the conditions of Lemma 2 hold for the function $f(s) = \mathcal{E}(s)$ in (3.1) and the r-sparse vector $s^\star$ in (C.11). The proof is by contradiction. Suppose that condition (1) of Lemma 2 is not satisfied, i.e. there exists an index $i \in \{1, 2, \dots, m\}$ such that $\dim(\{i\} \cup \operatorname{supp}(s^\star)) \leq r$, but the corresponding partial derivative

$$\frac{\partial \mathcal{E}(s^\star)}{\partial s_i} \triangleq \frac{\partial \mathcal{E}(s)}{\partial s_i}\bigg|_{s=s^\star} \tag{C.12}$$

is not zero; without loss of generality, assume that this partial derivative is positive:

$$\frac{\partial \mathcal{E}(s^\star)}{\partial s_i} > 0. \tag{C.13}$$

By the definitions of the partial derivative and limit, for the real number $\frac{1}{2} \frac{\partial \mathcal{E}(s^\star)}{\partial s_i}$, there exists a positive number $\delta > 0$ such that, for all $\epsilon \in (0, \delta)$, the vector $s_\epsilon = s^\star - \epsilon\, e_i$ satisfies

$$\left| \frac{\mathcal{E}(s^\star - \epsilon\, e_i) - \mathcal{E}(s^\star)}{-\epsilon} - \frac{\partial \mathcal{E}(s^\star)}{\partial s_i} \right| < \tfrac{1}{2} \frac{\partial \mathcal{E}(s^\star)}{\partial s_i} \tag{C.14a}$$

and, therefore,

$$\mathcal{E}(s^\star - \epsilon\, e_i) < \mathcal{E}(s^\star) - \tfrac{1}{2}\, \epsilon\, \frac{\partial \mathcal{E}(s^\star)}{\partial s_i}. \tag{C.14b}$$

Now, compute [see (3.2c)]

$$\mathcal{H}(s^\star - \epsilon\, e_i \,|\, s^\star) - \mathcal{H}(s^\star \,|\, s^\star) \;=\; (s^\star - \epsilon\, e_i - s^\star)^T \left[I_m - H^T \left(HH^T\right)^{-1} H \right] (s^\star - \epsilon\, e_i - s^\star) \tag{C.15a}$$

$$\leq\; \|s^\star - \epsilon\, e_i - s^\star\|_{\ell_2}^2 = \epsilon^2 \tag{C.15b}$$

where (C.15b) follows by observing that $H^T \left(HH^T\right)^{-1} H$ is positive semidefinite. Therefore, we have [see (3.2a)]

$$\mathcal{Q}(s^\star - \epsilon\, e_i \,|\, s^\star) \;=\; \mathcal{E}(s^\star - \epsilon\, e_i) + \mathcal{H}(s^\star - \epsilon\, e_i \,|\, s^\star) \tag{C.16a}$$

$$<\; \mathcal{E}(s^\star) - \tfrac{1}{2}\, \epsilon\, \frac{\partial \mathcal{E}(s^\star)}{\partial s_i} + \mathcal{H}(s^\star \,|\, s^\star) + \epsilon^2 \tag{C.16b}$$

$$=\; \mathcal{Q}(s^\star \,|\, s^\star) - \left(\tfrac{1}{2} \frac{\partial \mathcal{E}(s^\star)}{\partial s_i} - \epsilon \right) \epsilon \tag{C.16c}$$

where (C.16b) follows from (C.14b) and (C.15). Note that the vector $s_\epsilon = s^\star - \epsilon\, e_i$ is r-sparse. For any

$$\epsilon \in \left(0, \min \left\{ \delta, \tfrac{1}{2} \frac{\partial \mathcal{E}(s^\star)}{\partial s_i} \right\} \right) \tag{C.17}$$

we have $\mathcal{Q}(s_\epsilon \,|\, s^\star) < \mathcal{Q}(s^\star \,|\, s^\star)$, which contradicts (C.11). Hence, the condition (1) of Lemma 2 holds.

The condition (2) of Lemma 2 holds because, for any $s \in \mathcal{R}^m$, the Hessian of $\mathcal{E}(s)$ is

$$\frac{\partial^2 \mathcal{E}(s)}{\partial s\, \partial s^T} = 2\, H^T (HH^T)^{-1} H \tag{C.18}$$

which is clearly a positive semidefinite matrix.

Since the conditions of Lemma 2 hold for the function $f(s) = \mathcal{E}(s)$ in (3.1) and fixed point $s^\star$, we apply Lemma 2 and conclude that $s^\star$ is an r-local minimum point of $\mathcal{E}(s)$. Consequently, $s^\star$ is an r-local maximum point of the concentrated marginal likelihood function (2.6), which follows from the fact that (2.6) is a monotonically decreasing function of $\mathcal{E}(s) = N\, \widehat{\sigma}^2(s)$, see also (2.5). $\qquad\square$

APPENDIX D
PROOFS OF LEMMA 3, LEMMA 4, THEOREM 3 AND THEOREM 4

We first show Lemma 3.

Proof of Lemma 3: Equation (4.2a) in part (a) follows by noting that $H\boldsymbol{s} = H_A \boldsymbol{s}_A$, and (4.2b) in part (a) follows by using (4.2a) and the Rayleigh quotient property [35, Theorem 21.5.6], respectively:

$$
\begin{aligned}
\rho_{r,\min}(H) &= \min_{\boldsymbol{s} \in \mathcal{S}_r \setminus \mathbf{0}_{m \times 1}} \rho_r(\boldsymbol{s}, H) \\
&= \min_{A \subseteq \{1,2,\ldots,m\}, \dim(A)=r} \left[\min_{\boldsymbol{s}_A \in \mathcal{R}^r \setminus \mathbf{0}_{r \times 1}} \frac{\boldsymbol{s}_A^T H_A^T (HH^T)^{-1} H_A \boldsymbol{s}_A}{\|\boldsymbol{s}_A\|_{\ell_2}^2} \right] \\
&= \min_{A \subseteq \{1,2,\ldots,m\}, \dim(A)=r} \lambda_{\min}\left(H_A^T (HH^T)^{-1} H_A \right)
\end{aligned}
\tag{D.1}
$$

Part (b) holds because the row spaces of H and GH coincide.

The inequalities (4.4) in part (c) follow by applying the Rayleigh-quotient property and the fact that the projection matrix $H^T (HH^T)^{-1} H$ only has eigenvalues 0 and 1. When $r > N$, $H_A^T (HH^T)^{-1} H_A$ is not a full-rank matrix for any index set A with dimension r; consequently, $\rho_{r,\min}(H) = 0$ follows by using (4.2b) in part (a) of this lemma. When $N = m$, r-SSQ in (4.1a) is equal to one for any $0 < r \le m$, and, therefore, $\rho_{r,\min}(H) = 1$.

In part (d), we first show that (4.5) implies (4.6). When $\mathrm{spark}(H) > r$, the matrix $H_A^T(HH^T)^{-1}H_A$ is positive definite for any $A \subset \{1, 2, \ldots, m\}$ with $\dim(A) = r$, and (4.6) follows by using (4.2b) in part (a) of this lemma. We now show the 'only if' direction of part (d) by contradiction. Suppose that (4.6) holds but $\mathrm{spark}(H) \le r$. By the definition of spark, there exists an index set A with $\dim(A) = r$ such that the columns of H_A are linearly dependent. Therefore, the minimum eigenvalue of the matrix $H_A^T(HH^T)^{-1}H_A$ is zero and (4.2b) implies that $\rho_{r,\min}(H) = 0$, which leads to contradiction.

Finally, part (e) follows because the r_1-sparse vector that minimizes $\rho_{r_1}(\boldsymbol{s}, H)$ is also r_2-sparse and does not minimize $\rho_{r_2}(\boldsymbol{s}, H)$ in general. $\qquad\square$

Now, we prove Lemma 4.

Proof of Lemma 4: By part (d) of Lemma 3, the condition (4.10) holds if and only if $\mathrm{spark}(H) > 2r^\diamond$, which is exactly the condition required in [8, Theorem 2] to develop the same claim about the uniqueness of the (P_0) problem. This concludes the proof. $\qquad\square$

The proof of Theorem 3 is shown as follows.

Proof of Theorem 3: Let $\boldsymbol{s}^{(p)}$ be the estimate of $\boldsymbol{s}$ obtained in *Iteration* p of our ECME iteration. We

assume $s^{(p)} \neq s^\diamond$ without loss of generality; otherwise, the claim follows immediately. Now,

$$\|s^\diamond - s^{(p+1)}\|_{\ell_2}^2 = \frac{1}{\rho_{2r}(s^\diamond - s^{(p+1)}, H)} \|H^T (HH^T)^{-1} H(s^\diamond - s^{(p+1)})\|_{\ell_2}^2 \tag{D.2a}$$

$$\leq \frac{1}{\rho_{2r,\min}(H)} \mathcal{E}(s^{(p+1)}) \tag{D.2b}$$

$$= \frac{1}{\rho_{2r,\min}(H)} \left[\mathcal{Q}(s^{(p+1)} \,|\, s^{(p)}) - \mathcal{H}(s^{(p+1)} \,|\, s^{(p)}) \right] \tag{D.2c}$$

$$\leq \frac{1}{\rho_{2r,\min}(H)} \mathcal{Q}(s^{(p+1)} \,|\, s^{(p)}) \tag{D.2d}$$

$$\leq \frac{1}{\rho_{2r,\min}(H)} \mathcal{Q}(s^\diamond \,|\, s^{(p)}) \tag{D.2e}$$

$$= \frac{1}{\rho_{2r,\min}(H)} \|s^\diamond - s^{(p)} - H^T(HH^T)^{-1}H\,(s^\diamond - s^{(p)})\|_{\ell_2}^2 \tag{D.2f}$$

$$= \frac{1}{\rho_{2r,\min}(H)} (s^\diamond - s^{(p)})^T[I_m - H^T(HH^T)^{-1}H]\,(s^\diamond - s^{(p)}) \tag{D.2g}$$

$$= \frac{1}{\rho_{2r,\min}(H)} \left[\|s^\diamond - s^{(p)}\|_{\ell_2}^2 - \rho_{2r}(s^\diamond - s^{(p)}, H)\,\|s^\diamond - s^{(p)}\|_{\ell_2}^2\right] \tag{D.2h}$$

$$\leq \zeta(H)\,\|s^\diamond - s^{(p)}\|_{\ell_2}^2 \tag{D.2i}$$

where

$$\zeta(H) \triangleq \frac{1 - \rho_{2r,\min}(H)}{\rho_{2r,\min}(H)} \tag{D.3}$$

and (D.2a) follows from the definition (4.1a) and the fact that $s^\diamond - s^{(p+1)}$ is at most $2r$-sparse since $s^\diamond$ and $s^{(p+1)}$ are r-sparse; (D.2b) is due to the definitions (3.1) and (4.1b), (D.2c) results from the identity (3.2a); (D.2d) holds because $\mathcal{H}(s^{(p+1)} \,|\, s^{(p)})$ is nonnegative [see (3.2c)]; (D.2e) follows due to the M step of the ECME algorithm (2.7b); (D.2f) uses the definition (3.2b) and the condition (4.14a); (D.2g)–(D.2i) follow by expanding (D.2f) and using the definitions (4.1a), (4.1b), and (D.3), respectively.

We now apply the condition (4.14b) and conclude that $\zeta(H)$ in (D.3) is nonnegative and smaller than one:

$$0 \leq \zeta(H) = \frac{1 - \rho_{2r,\min}(H)}{\rho_{2r,\min}(H)} < 1. \tag{D.4}$$

Therefore, the sequence $\|s^\diamond - s^{(p)}\|_{\ell_2}^2$ monotonically shrinks to zero and the claim follows. $\qquad\square$

Finally, we prove Theorem 4.

Proof of Theorem 4: Denote by $s^{(p)}$ the sparse signal estimate in *Iteration* p of our ECME iteration.

Now, for $p \geq 0$

$$\|\boldsymbol{s}^{(p+1)} - \boldsymbol{s}_r^\diamond\|_{\ell_2}$$

$$\leq \frac{\|H^T (HH^T)^{-1} H (\boldsymbol{s}^{(p+1)} - \boldsymbol{s}_r^\diamond)\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5a}$$

$$= \frac{\|H^T (HH^T)^{-1} (\boldsymbol{y} - H\boldsymbol{s}^{(p+1)}) + H^T (HH^T)^{-1} H (\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond) - H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5b}$$

$$\leq \frac{\sqrt{\mathcal{E}(\boldsymbol{s}^{(p+1)})} + \|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond\|_{\ell_2} + \|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5c}$$

$$\leq \frac{\sqrt{\mathcal{Q}(\boldsymbol{s}_r^\diamond \,|\, \boldsymbol{s}^{(p)})} + \|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond\|_{\ell_2} + \|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5d}$$

$$\leq \frac{\|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^{(p)} - H^T(HH^T)^{-1}[H\boldsymbol{s}_r^\diamond + H(\boldsymbol{s}^\diamond - \boldsymbol{s}_r^\diamond) + \boldsymbol{n} - H\,\boldsymbol{s}^{(p)}]\|_{\ell_2} + \|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond\|_{\ell_2} + \|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}}$$

$$\tag{D.5e}$$

$$\leq \frac{\|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^{(p)} - H^T (HH^T)^{-1} H (\boldsymbol{s}_r^\diamond - \boldsymbol{s}^{(p)})\|_{\ell_2} + 2\|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond\|_{\ell_2} + 2\|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5f}$$

$$\leq [\zeta(H)]^{1/2} \|\boldsymbol{s}^{(p)} - \boldsymbol{s}_r^\diamond\|_{\ell_2} + \frac{2\|\boldsymbol{s}_r^\diamond - \boldsymbol{s}^\diamond\|_{\ell_2} + 2\|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}}{\sqrt{\rho_{2r,\min}(H)}} \tag{D.5g}$$

where (D.5a) follows from the definition (4.1b) and the fact that $\boldsymbol{s}^{(p+1)} - \boldsymbol{s}_r^\diamond$ is $2r$-sparse; (D.5b) follows by using (4.15a); in (D.5c), we use the triangle inequality ($\|\boldsymbol{a} + \boldsymbol{b}\|_{\ell_2} \leq \|\boldsymbol{a}\|_{\ell_2} + \|\boldsymbol{b}\|_{\ell_2}$), definition (3.1), and the fact that the eigenvalues of $H^T (HH^T)^{-1} H$ are 0 and 1; (D.5d) follows along the same lines as (D.2b)–(D.2e) with $\boldsymbol{s}^\diamond$ replaced by $\boldsymbol{s}_r^\diamond$; (D.5e) follows from (3.2b) and (4.15a); (D.5f) holds due to the the triangle inequality and the fact that the eigenvalues of $H^T (HH^T)^{-1} H$ are 0 and 1; finally, (D.5g) follows from the same lines as (D.2f)–(D.2i) with $\boldsymbol{s}^\diamond$ replaced by $\boldsymbol{s}_r^\diamond$.

From (D.5g), we can see by induction that, for $p \geq 1$

$$\|\boldsymbol{s}^{(p)} - \boldsymbol{s}_r^\diamond\|_{\ell_2} \leq [\zeta(H)]^{p/2} \|\boldsymbol{s}^{(0)} - \boldsymbol{s}_r^\diamond\|_{\ell_2} + 2\frac{\sum_{i=0}^{p-1}[\zeta(H)]^{i/2}}{\sqrt{\rho_{2r,\min}(H)}} \left[\|\boldsymbol{s}^\diamond - \boldsymbol{s}_r^\diamond\|_{\ell_2} + \|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}\right]$$

$$= [\zeta(H)]^{p/2} \|\boldsymbol{s}^{(0)} - \boldsymbol{s}_r^\diamond\|_{\ell_2} + 2\frac{1}{\sqrt{\rho_{2r,\min}(H)}} \frac{1 - [\zeta(H)]^{p/2}}{1 - [\zeta(H)]^{1/2}} \left[\|\boldsymbol{s}^\diamond - \boldsymbol{s}_r^\diamond\|_{\ell_2} + \|H^T (HH^T)^{-1}\boldsymbol{n}\|_{\ell_2}\right] \tag{D.6}$$

where $\boldsymbol{s}^{(0)}$ is the initial signal estimate. Since the condition (4.15c) implies that $\zeta(H)$ in (D.3) is nonnegative and smaller than one [see (D.4)], we have

$$\lim_{p \nearrow +\infty} [\zeta(H)]^{p/2} = 0 \tag{D.7}$$

the first term in (D.6) disappears (i.e. the effect of the initial signal estimate washes out), and the claim follows. $\square$

APPENDIX E
PROOF OF THEOREM 5

Proof of Theorem 5: When conditions (1) and (2) of Theorem 5 hold, we have $\mathrm{spark}(H) = N+1 > 2\,r^\diamond$ and the condition of [8, Theorem 2] is satisfied. Therefore, $\boldsymbol{s}^\diamond$ is the unique solution of the (P_0) problem, according to [8, Theorem 2]. We now consider the USS function under different sparsity level r.

For $r = r^\diamond$, the ML estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(r^\diamond) = \big(\widehat{\boldsymbol{s}}_{\mathrm{ML}}(r^\diamond), \widehat{\sigma}^2_{\mathrm{ML}}(r^\diamond)\big) = (\boldsymbol{s}^\diamond, 0)$ and unique, since it leads to infinite likelihood function (2.4a) and no other $\boldsymbol{\theta}$ yields infinite likelihood, due to the fact that $\boldsymbol{s}^\diamond$ is the unique solution of the (P_0) problem. Furthermore, since (6.2) holds, we have $N - r^\diamond - 2 > 0$ and therefore $\mathrm{USS}(r)$ is infinite as well. Note that

$$\mathrm{USS}(r) = \widetilde{\mathrm{USS}}\big(r, \widehat{\sigma}^2_{\mathrm{ML}}(r)\big) \tag{E.1}$$

where

$$\widetilde{\mathrm{USS}}(r, \sigma^2) = -\tfrac{1}{2}\, r\, \ln\Big(\frac{N}{m}\Big) - \tfrac{1}{2}\,(N - r - 2)\, \ln\Big(\frac{\sigma^2}{\boldsymbol{y}^T\,(H\,H^T)^{-1}\,\boldsymbol{y}/N}\Big). \tag{E.2}$$

Now,

$$\lim_{\sigma^2 \searrow 0} \frac{\widetilde{\mathrm{USS}}(r^\diamond, \sigma^2)}{\ln(1/\sigma^2)} = \tfrac{1}{2}\,(N - r^\diamond - 2) > 0 \tag{E.3}$$

specifying the rate of growth to infinity of $\widetilde{\mathrm{USS}}(r^\diamond, \sigma^2)$ as σ^2 approaches the ML estimate $\widehat{\sigma}^2_{\mathrm{ML}}(r^\diamond) = 0$.

For $r < r^\diamond$, $\boldsymbol{y} \neq H\,\boldsymbol{s}$ for any r-sparse vector $\boldsymbol{s}$; consequently, $\sigma^2_{\mathrm{ML}}(r) > 0$ and $\mathrm{USS}(r)$ is finite.

For $r > r^\diamond$, the ML estimate of σ^2 must be $\widehat{\sigma}^2_{\mathrm{ML}}(r) = 0$, which leads to infinite likelihood. However, in this case,

$$\lim_{\sigma^2 \searrow 0} \frac{\widetilde{\mathrm{USS}}(r, \sigma^2)}{\ln(1/\sigma^2)} = \tfrac{1}{2}\,(N - r - 2) < \tfrac{1}{2}\,(N - r^\diamond - 2). \tag{E.4}$$

Therefore, if $r \geq N - 2$, $\mathrm{USS}(r)$ is either finite or goes to negative infinity. For $r^\diamond < r < N - 2$, $\mathrm{USS}(r)$ is infinitely large, but the rate at which $\widetilde{\mathrm{USS}}(r, \sigma^2)$ grows to infinity as σ^2 approaches the ML estimate $\widehat{\sigma}^2_{\mathrm{ML}}(r) = 0$ is smaller than that specified by (E.3).

The claim follows by combining the above conclusions. □

REFERENCES

[1] I.F. Gorodnitsky and B.D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, pp. 600–616, Mar. 1997.

[2] E.J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, Feb. 2006.

[3] W.U. Bajwa, J.D. Haupt, A.M. Sayeed, and R.D. Nowak, "Joint source-channel communication for distributed estimation in sensor networks," *IEEE Trans. Inform. Theory*, vol. 53, pp. 3629–3653, Oct. 2007.

[4] M. Lustig, D. Donoho, and J.M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, Dec. 2007.

[5] R.G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, pp. 118–121, Jul. 2007.

[6] *IEEE Signal Processing Mag. Special Issue on Sensing, Sampling, and Compression*, Mar. 2008.

[7] E.J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[8] A.M. Bruckstein, D.L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, pp. 34–81, Mar. 2009.

[9] B.K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[10] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[11] E.J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate information," *Commun. Pure and Applied Mathematics*, vol. 59, pp. 1207–1233, Aug. 2006.

[12] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, pp. 2313–2351, Dec. 2007.

[13] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems," *IEEE J. Select. Areas Signal Processing*, vol. 1, pp. 586–597, Dec. 2007.

[14] S. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[15] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation", *IEEE Trans. Inform. Theory*, vol. 50, pp. 2231–2242, Oct. 2004.

[16] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[17] D. Needell and J.A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, vol. 26, pp. 301–321, May 2009.

[18] K.K. Herrity, A.C. Gilbert, and J.A. Tropp, "Sparse approximation via iterative thresholding," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* Toulouse, France, May 2006, pp. 624–627.

[19] T. Blumensath and M.E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, pp. 629–654, Dec. 2008.

[20] T. Blumensath and M.E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comp. Harmonic Anal.*, vol. 27, pp. 265–274, Nov. 2009.

[21] T. Blumensath and M.E. Davies, "Normalized iterative hard thresholding; guaranteed stability and performance," *IEEE J. Select. Areas Signal Processing*, vol. 4, pp. 298–309, Apr. 2010.

[22] D.P. Wipf and B.D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, pp. 2153–2164, Aug. 2004.

[23] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, pp. 2346–2356, Jun. 2008.

[24] K. Qiu and A. Dogandžić, "Variance-component based sparse signal reconstruction and model selection," *IEEE Trans. Signal Processing*, vol. 58, pp. 2935–2952, Jun. 2010.

[25] A. Maleki and D.L. Donoho, "Optimally tuned iterative thresholding algorithms for compressed sensing," *IEEE J. Select. Areas Signal Processing*, vol. 4, pp. 330–341, Apr. 2010.

[26] A. Dogandžić and K. Qiu, "Automatic hard thresholding for sparse signal reconstruction from NDE measurements," in *Rev. Progress Quantitative Nondestructive Evaluation*, D.O. Thompson and D.E. Chimenti (Eds.), Melville NY: Amer. Inst. Phys., vol. 29, 2010, pp. 806–813.

[27] K. Qiu and A. Dogandžić, "Double overrelaxation thresholding methods for sparse signal reconstruction," in *Proc. 44th Annu. Conf. Inform. Sci. Syst.*, Princeton, NJ, Mar. 2010.

[28] Y.X. He and C.H. Liu, "The dynamic ECME algorithm," Dept. Statistics, Purdue Univ., Tech. Report, 2009.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc., Ser. B*, vol. 39, pp. 1–38, July 1977.

[30] C.F.J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, pp. 95–103, Mar. 1983.

[31] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd. ed., New York: Wiley, 2008.

[32] C.H. Liu and D.B. Rubin, "The ECME algorithm: A simple extension of EM and ECM with fast monotone convergence," *Biometrika*, vol. 81, pp. 633–648, Dec. 1994.

[33] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[34] R.A. Thisted, *Elements of Statistical Computing: Numerical Computation*, Chapman & Hall, 1988.

[35] D.A. Harville, *Matrix Algebra From a Statistician's Perspective*, New York: Springer-Verlag, 1997.

[36] S.J. Colley, *Vector Calculus*, 3rd. ed., Upper Saddle River, NJ: Prentice Hall, 2006.

[37] D.L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition", *IEEE Trans. Inform. Theory*, vol. 47, pp. 2845–2862, Nov. 2001.

[38] D.L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization", *Proc. Nat. Acad. Sci. USA*, vol. 100, pp. 2197–2202, Mar. 2003.

[39] E.J. Candès and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[40] A.V. Oppenheim and R.W. Schafer *Discrete-time Signal Processing*, 3rd ed., Upper Saddle River, NJ: Prentice Hall, 2010.

[41] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia: SIAM, 1992.

[42] T.T. Do, T.D. Tran, and L. Gan, "Compressive sampling with structurally random matrices," in *Proc. Int. Conf. Acoust., Speech, Signal Processing,* Las Vegas, NV, pp. 3369–3372, Apr. 2008.